Une méthode stable de bloc-diagonalisation de matrices: Application au calcul de portrait spectral

Pierre-François Lavallée, Miloud Sadkane

▶ **To cite this version:**

Pierre-François Lavallée, Miloud Sadkane. Une méthode stable de bloc-diagonalisation de matrices: Application au calcul de portrait spectral. [Rapport de recherche] RR-3141, INRIA. 1997. inria-00073548

HAL Id: inria-00073548
https://hal.inria.fr/inria-00073548

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Une méthode stable de bloc-diagonalisation de matrices: Application au calcul de portrait spectral

Pierre-François Lavallée et Miloud Sadkane

N° 3141

Mars 1997

THÈME 4

ISSN 0249-6399

INRIA
RENNES

Une méthode stable de bloc-diagonalisation de matrices: Application au calcul de portrait spectral

Pierre-François Lavallée et Miloud Sadkane

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet ALADIN

Rapport de recherche n° 3141 — Mars 1997 — 57 pages

Résumé : Ce rapport présente une manière originale et peu coûteuse d'approximer le portrait spectral d'une matrice A. L'idée directrice est de décomposer l'espace $\mathbb{C}^n$ en la somme directe de r sous-espaces stables $M_i, \quad i = 1, \ldots, r$. Ces sous-espaces sont déterminés de façon à maximiser l'angle canonique minimal entre M_i et $\cup_{j \neq i} M_j$. On peut montrer que cela revient à réduire la matrice A en une matrice diagonale par bloc $D = S^{-1}AS$, tout en imposant au conditionnement $K(S) = \|S\|\|S^{-1}\|$ de la matrice de passage S de rester aussi petit que possible. Dans le cas où une telle décomposition est effectivement déterminée avec un $K(S)$ raisonnable, on montre qu'alors le portrait spectral de D est une bonne approximation de celui de A. La prise en compte de la structure diagonale par blocs de la matrice D nous permet de réduire les coûts de calcul de façon significative. Ces résultats sont illustrés sur plusieurs exemples.

Mots-clé : valeur propre, ϵ-spectre, portrait spectral, conditionnement, Schur, Sylvester, angles canoniques de sous-espaces.

(Abstract: pto)

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 02 99 84 71 00 – Télécopie : (33) 02 99 84 71 71

A stable block diagonalization method : Application to the spectral portait of matrices

Abstract: This report deals with an original and a cheap way of approximating the spectral portrait of a matrix A. The main idea is to decompose the n-dimensional space $\mathbb{C}^n$ into a direct sum of r stable subspaces $M_i, \quad i = 1, \ldots, r$. These subspaces are determined in a way such that the sine between M_i and $\cup_{j \neq i} M_j$ is maximized. It can be schown that this amounts to reducing the matrix A into a block-diagonal matrix $D = S^{-1}AS$, where the condition number $K(S) = \|S\|\|S^{-1}\|$ remains small. In the case where such a stable decomposition is found, with a reasonable $K(S)$, we show that the spectral portrait of the matrix D is a good approximation of the one of A. Tacking into account the block-diagonal structure of D allows us to reduce the computation of the spectral portrait significantly. The results are illustrated by several examples.

Key-words: eigenvalue, ϵ-spectrum, spectral portrait, condition number, Schur, Sylvester, canonical angles between subspaces.

1 Introduction

Soit A une matrice d'ordre n, à éléments complexes, et désignons par $\| \cdot \|$ la norme Euclidienne ou sa norme matricielle induite. La notion du ϵ-spectre introduite indépendamment par Trefethen [13] et Godunov [5] généralise la notion de valeur propre, en ce sens qu'au lieu de représenter une valeur propre par une valeur approchée, on considère un voisinage de cette valeur défini par un seuil $\epsilon \geq 0$. Le ϵ-spectre de A, noté dans la suite $\Lambda_\epsilon(A)$, est l'ensemble des valeurs propres de toutes les matrices perturbées de la forme $A + \Delta$ avec $\|\Delta\| \leq \epsilon\|A\|$.

$$\begin{aligned} \Lambda_\epsilon(A) &= \{z \in \mathbb{C} \ / \ z \text{ est valeur propre de } A+\Delta,\ \|\Delta\| \leq \epsilon\|A\|\} \quad (1) \\ &= \{z \in \mathbb{C} \ / \ \det(A+\Delta-zI) = 0,\ \|\Delta\| \leq \epsilon\|A\|\}\ . \quad (2) \end{aligned}$$

Il est connu [6] que, contrairement à la plus petite valeur propre, la plus petite valeur singulière d'une matrice donne une indication sur la proximité d'une matrice singulière. Si $z \in \mathbb{C}$ alors

$$\sigma_{min}(A - zI) = \min\{\|\Delta\| \ / \ z \text{ est valeur propre de } A + \Delta\}. \quad (3)$$

Ainsi on obtient la définition équivalente

$$\Lambda_\epsilon(A) = \{z \in \mathbb{C} \ / \ \sigma_{min}(A - zI) \leq \epsilon\|A\|\}\ , \quad (4)$$

qui peut encore s'écrire sous la forme

$$\Lambda_\epsilon(A) = \{z \in \mathbb{C} \ / \ \|(A - zI)^{-1}\|_2 \geq \frac{1}{\epsilon\|A\|}\}\ , \quad (5)$$

avec la convention $\|(A - zI)^{-1}\|_2 = \infty$ si z est une valeur propre de A. Autrement dit, $\Lambda_0(A)$, le 0-spectre de A n'est rien d'autre que l'ensemble des valeurs propres de A et c'est ainsi qu'on le notera dans la suite.
Si la matrice A est normale, on a :

$$\|(A - zI)^{-1}\|_2 = \frac{1}{\text{dist}(z, \Lambda_0(A))}, \quad (6)$$

où $\text{dist}(z, S)$ est la distance usuelle entre z et l'ensemble S. Ainsi, à ϵ fixé, $\Lambda_\epsilon(A)$ est en fait l'ensemble des disques de rayon $\epsilon\|A\|$ et de centre les points constituant le spectre de A. Dans ce cas, $\Lambda_\epsilon(A)$ est entièrement déterminé par la donnée d'un ϵ et la connaissance du spectre de A.
Dans le cas où la matrice A est non normale, l'égalité (6) devient :

$$\|(A - zI)^{-1}\|_2 \geq \frac{1}{dist(z, \Lambda_0(A))}. \quad (7)$$

Dans ce cas, on ne peut plus déduire explicitement $\Lambda_\epsilon(A)$, simplement à partir de la connaissance du spectre de A et de ϵ. Le fait que l'on puisse avoir $\|(A - zI)^{-1}\|_2$ de l'ordre de 10^{10} ou 10^{20}, même pour des z éloignés du spectre de A, montre que la surface représentant $\Lambda_\epsilon(A)$ pourra être très étalée dans le plan complexe. C'est la représentation de ces taches dans le plan complexe que l'on appelle le portrait spectral de la matrice A. Une des manières de visualiser ce portrait spectral est la représentation de la fonction $sp_A(z) = \log_{10}(\sigma_{min}(zI - A))$ par l'intermédiaire de courbes de niveaux dans le plan complexe. Notons que $\sigma_{min}(zI - A) = 1/\|R(A, z)\|$ où $R(A, z) = (A - zI)^{-1}$ est la résolvante de la matrice au point z. Ainsi le portrait spectral permet de visualiser le comportement de la norme de la résolvante. Une façon classique de procéder pour calculer sp_A est d'utiliser l'algorithme de décomposition en valeurs singulières [6] pour chaque point z d'une grille qui discrétise une partie du plan complexe. Cette approche est acceptable mais requiert un coût de calcul très élevé.
Nous proposons une autre méthode qui consiste dans un premier temps à réduire la matrice A en une matrice diagonale par bloc de la forme

$$A = S \operatorname{diag}(D_1, D_2, \ldots, D_q)\, S^{-1},\ 1 \le q \le n, \tag{8}$$

avec la contrainte que le conditionnement $\kappa(S) = \|S\|\|S^{-1}\|$ reste inférieur à une borne donnée par l'utilisateur. De cette façon, on espère que les propriétés spectrales de la matrice bloc-diagonale D vont être proches de celles de A. L'algorithme de bloc-diagonalisation est inspiré de celui proposé en [1]. On commence par factoriser A sous une forme de Schur $A = QTQ^*$. Ensuite, la matrice triangulaire supérieure T est bloc-diagonalisée, de telle sorte que chaque bloc diagonal contienne des valeurs propres correspondant à des vecteurs propres qui sont proches (en un sens que l'on précisera en Section 2). Le portrait spectral de la matrice A est alors approché par celui des matrices $D_i,\ i = 1, \ldots, q$.
L'avantage de cette méthode par rapport à la méthode classique [5, 13] est sa simplicité relative et sa rapidité. En effet, le portrait spectral de chacune des petites matrices $D_i,\ i = 1, \ldots, q$, est non seulement peu coûteux en calcul, mais peut en plus être fait en parallèle.
Le plan de ce rapport est le suivant. En section 2, nous décrivons et justifions notre algorithme de bloc-diagonalisation. Nous comparons différents critères utilisés pour le choix des blocs. Dans la section 3, nous établissons différentes relations entre le portrait spectral de la matrice A et celui des matrices $D_1, \ldots, D_q$ de l'égalité (8). Nous discutons aussi la possibilité d'utiliser le champ des valeurs de ces matrices. La section 4 est consacrée à l'analyse des coûts de calcul et des résultats numériques.

2 Algorithme de bloc-diagonalisation

Une décomposition de la forme (8) a de nombreuses applications. Une des plus connues est le calcul des puissances de A, qui, lorsque les blocs D_i sont de "petites" tailles, peut se faire facilement. En effet, on a alors

$$A^k = S \operatorname{diag}(D_1^k, D_2^k, \ldots, D_q^k)\, S^{-1}\ .$$

Dans notre cas nous allons faire une utilisation novatrice de cette décomposition, puisque l'on va s'en servir pour approcher le portrait spectral de la matrice A.
Rappelons quelques propriétés de cette décomposition. Si S est partitionnée en q blocs colonnes compatibles avec la décomposition en bloc de D, (i.e. $S = [S_1|\dots|S_q]$), alors les colonnes de S_i forment une base d'un sous-espace S^i, invariant par A (i.e. $AS^i \subset S^i$) et D_i est la représentation de A dans cette base. Le projecteur spectral associé est S_iY_i, où Y_i est formé des lignes correspondantes de S^{-1}.
Il y a des limitations théoriques et pratiques sur la taille des blocs D_i. Théoriquement, il ne peuvent pas être plus petit que les blocs correspondants issus de la décomposition de Jordan de A. En pratique, ils doivent être plus grand. Les problèmes numériques associés à la décomposition (8) ont été étudiés en détail dans [7]. On retiendra que la principale difficulté provient du fait que la forme de Jordan d'une matrice n'est pas stable numériquement (voir [10]). De petites perturbations de la matrice A peuvent en effet engendrer la fusion ou la séparation de plusieurs blocs. Ainsi, si on sépare deux blocs "proches" l'un de l'autre, alors les colonnes de S tendent à devenir linéairement dépendentes. Autrement dit, le conditionnement de S: $K(S) = \|S\|\|S^{-1}\|$ sera grand. Dans ce cas, il sera impossible de calculer S^{-1} ou de résoudre un système linéaire du type $Sx = b$ avec précision [11, p.170]. Une étude approfondie de ces phénomènes a été menée par Demmel dans [2], dans le cas où l'on utilise une factorisation de A du type (8), pour calculer $f(A)$ où f est une fonction analytique de A.

2.1 Étude des différentes étapes de la bloc-diagonalisation de A

L'idée directrice est d'utiliser la décomposition de Schur pour réduire A en une matrice triangulaire supérieure

$$Q^*AQ = T = \begin{pmatrix} T_{11} & T_{12} & \dots & T_{1q} \\ & T_{22} & \dots & T_{2q} \\ & & \ddots & \vdots \\ & & & T_{qq} \end{pmatrix} \tag{9}$$

et ensuite de bloc-diagonaliser T à l'aide du théorème suivant (voir [6, p.338])

Théorème 1
Supposons que la décomposition (9) soit telle que les matrices T_{ii} et T_{jj} aient des spectres disjoints pour $i \neq j$. Alors, il existe une matrice non singulière X telle que

$$(QX)^{-1}A(QX) = diag(T_{11}, \dots, T_{qq}). \tag{10}$$

On recherche alors des conditions suffisantes et raisonnables pour que la factorisation issue du théorème précédent soit stable. En d'autres termes , on voudrait trouver une matrice S non singulière telle que $S^{-1}AS = \mathrm{diag}(D_{11}, \dots, D_{qq})$ avec la propiété "$\kappa(S)$ petit". Cette propriété est une conséquence directe du choix des caractéristiques des différents blocs (i.e. taille et nombre de blocs, éléments propres les constituant, etc...).

2.1.1 Détermination des blocs et de la matrice S

Premier critère de choix des blocs
Une des raisons qui nous a conduit à utiliser la factorisation bloc-diagonale de la matrice A vient de la définition même du portrait spectral : il peut être divisé en plusieurs "taches", chacune d'elle contenant des valeurs propres d'une matrice $A + E$, où la norme de E est inférieure à $\epsilon\|A\|$. D'où l'idée de regrouper dans un même bloc diagonal les valeurs propres qui appartiennent à la même "tache".
On définit ainsi un premier critère pour la détermination du choix des blocs : on regroupe dans un même bloc des valeurs propres qui sont "proches" les unes des autres. En d'autres termes, étant donné $\eta > 0$, on regroupe dans un même bloc les valeurs propres distantes de moins de η. Cette méthode a donc été implantée et testée sur plusieurs exemples numériques (voir paragraphe 2.3). Ceux-ci ont très vite révélés les limites de ce choix. Le conditionnement $K(S)$ de la matrice S issue d'une telle décomposition croît beaucoup trop vite lorsque le nombre de blocs augmente.

On a donc été amené à choisir un autre critère pour la détermination du choix des blocs, nous permettant de limiter au mieux la croissance de $K(S)$.

Nouveau critère de choix des blocs
C'est grâce aux résultats suivants, dus à Demmel [2], que l'on a pu déterminer ce nouveau critère. Rappelons tout d'abord quelques uns de ces résultats issus de [2].

Notations:

- $S = [S_1|\ldots|S_q]$ une partition de S en q blocs colonnes,
- S^i le sous-espace engendré par les colonnes de S_i,
- $\mathcal{V}_{ij} = \mathcal{V}(S^i, S^j)$ l'angle entre S^i et S^j,
- $\mathcal{V}_i = \mathcal{V}(S^i, \mathrm{eng}_{j\neq i}\{S^j\})$ angle entre S^i et le sous-espace engendré par $\bigcup_{j\neq i} S^j$.
- $1/\sin(.) \equiv \csc(.)$

Si $S_i^* S_i = I$ pour $i = 1, \ldots q$, (i.e. les colonnes de S_i forment une base orthonormale de S^i), alors

$$\begin{aligned} \mathcal{V}_{ij} &= \min\{arccos|u^*v|/u \in S^i, v \in S^j, \|u\| = \|v\| = 1\} \\ &= arccos(\sup_{x,y} |y^* S_i^* S_j x|) \\ &= arccos\|S_i^* S_j\|. \end{aligned}$$

Théorème 2
Supposons que $S_i^* S_i = I$ *pour* $i = 1, \dots q$, *alors*

$$\kappa(S) \leq \sqrt{q}\, \kappa(S_{OPT}) \tag{11}$$

$$\max_i(\csc \mathcal{V}_i + \sqrt{\csc^2 \mathcal{V}_i - 1}) \leq \kappa(S) \leq \sqrt{q}\sqrt{\sum_{i=1}^{q} \csc^2 \mathcal{V}_i}\ , \tag{12}$$

où S_{OPT} *est une matrice* S *dont le conditionnement est optimal, c'est à dire le plus petit possible.*

Théorème 3
Dans le cas où $q = 2$, *soient* $T = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}$, R *une solution de l'équation de Sylvester* $T_{11}R - RT_{22} = T_{12}$ *et* $S = \begin{pmatrix} I & -R \\ 0 & I \end{pmatrix}$. *Alors, si on note* $P = \begin{pmatrix} I & R \\ 0 & 0 \end{pmatrix}$, *le projecteur associé à* S^1, *parallèlement à* S^2, *on a*

$$S^{-1}TS = diag(T_{11}, T_{22}),$$

et

$$\kappa(S) \geq \cot \mathcal{V}/2 \quad \text{avec } S = [S_1|S_2] \text{ et } \mathcal{V} = \mathcal{V}(S_1, S_2). \tag{13}$$

Si de plus $S_i^* S_i = I,\ i = 1, 2$, *alors*

$$\kappa(S) = \kappa(S_{OPT}) = \cot \mathcal{V}/2 = \sqrt{\|R\|^2 + 1} + \|R\| \tag{14}$$

$$1/\sin \mathcal{V} \equiv \csc \mathcal{V} = \|P\|\ . \tag{15}$$

Les théorèmes 2 et 3 sont à la base de la détermination d'un critère de sélection des différents blocs $D_i,\ i = 1, \dots, q$. Deux approches différentes sont alors possibles.

La première consiste à remarquer que lorsqu'on a deux blocs, pour rendre $K(S)$ petit, il est nécessaire d'avoir $\|R\|$ petit en vertu de l'égalité (14) du théorème 3. La détermination des blocs se fait de manière récurrente de la façon suivante :

- étant donnée une décomposition de la matrice T de la forme $\begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}$, on résout le système de Sylvester associé $T_{11}R - RT_{22} = T_{12}$,
- on calcule $\|R\|$,

- si cette norme est plus petite qu'une borne limite vl prédéfinie, on effectue l'élimination du terme T_{12} (nous verrons plus précisément au paragraphe 2.1.3 la démarche à suivre). T_{11} est alors le premier bloc, et on recommence récursivement avec le bloc T_{22} que l'on décompose à son tour en deux sous-blocs. Dans le cas où $\|R\| > vl$, ce qui impliquerait si on faisait l'élimination, un grand $K(S)$, ($K(S) > \sqrt{vl^2+1} + vl$), on augmente la dimension de T_{11} et on recommence au premier point.

C'est la méthode utilisée par Stewart dans [1] pour une matrice A à coefficients réels.

La seconde approche est une conséquence des constatations suivantes. Soit (λ_k, u_k) avec $\|u_k\| = 1,\ k = 1, \ldots, n$, les éléments propres de la matrice A. Supposons par exemple que λ_i et λ_j ne soient pas dans le même bloc et que

$$\cos \mathcal{V}(u_i, u_j) = |u_i^* u_j| \geq 1 - \eta, \quad 0 \leq \eta \ll 1 \tag{16}$$

alors

$$\operatorname{csc} \mathcal{V}(u_i, u_j) \geq (2\eta - \eta^2)^{-\frac{1}{2}}. \tag{17}$$

De (12) et (17) on déduit

$$\kappa(S) \geq \max_k \operatorname{csc} \mathcal{V}_k \geq \operatorname{csc} \mathcal{V}(u_i, u_j) \approx 1/\sqrt{2\eta}\ ,$$

ce qui revient à dire que $K(S)$ sera "grand" pour η proche de zéro. On en déduit une règle simple pour la détermination des blocs :
Pour que $K(S)$ reste raisonnablement "petit", il est nécessaire que deux vecteurs propres normés u_i et u_j correspondant respectivement aux deux valeurs propres λ_i et λ_j de deux blocs différents doivent être tels que $|u_i^* u_j| < 1 - \eta$ où η est un paramètre petit, fixé par l'utilisateur.

Détermination effective des blocs
La détermination effective des blocs peut se modéliser de la façon suivante. Soit $E = \{1, 2, \ldots, n\}$ et G le graphe non orienté défini sur E par :

$$(i, j) \in G \quad \Longleftrightarrow \quad |u_i^* u_j| \geq 1 - \eta$$

Alors les différents blocs correspondent aux différentes composantes connexes du graphe G. Le nombre de blocs pour η fixé, est égal au nombre de composantes connexes différentes du graphe G. C'est ce que l'on va appeler la décomposition maximale admissible pour un η donné.

Exemple d'une décomposition maximale admissible

- Soit T une matrice triangulaire supérieure, provenant de la décomposition de Schur de la forme $T = \begin{pmatrix} \lambda_1 & \dots & \dots & \dots & \dots \\ & \lambda_2 & \dots & \dots & \dots \\ & & \lambda_3 & \dots & \dots \\ & & & \lambda_4 & \dots \\ & & & & \lambda_5 \end{pmatrix}$.
- Soit $u_1, \dots, u_5$ les vecteurs propres associés normés à l'unité.
- Soit $\eta = 0.1$ donné.

Calculons la matrice $Test$ de terme générique $|u_i^* u_j|$, on obtient le résultat suivant :

$$Test = \begin{pmatrix} 1 & 0.2 & 0.95 & 0.8 & 0.37 \\ & 1 & 0.54 & 0.28 & 0.96 \\ & & 1 & 0.98 & 0.65 \\ & & & 1 & 0.1 \\ & & & & 1 \end{pmatrix}.$$

De la matrice $Test$, on déduit le graphe G :

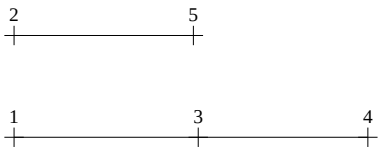

FIG. 1: *Graphe G*

On a deux composantes connexes distinctes, donc deux blocs $B_1 = \{\lambda_1, \lambda_3, \lambda_4\}$ et $B_2 = \{\lambda_2, \lambda_5\}$.

2.1.2 Réordonnancement des valeurs propres sur la diagonale

On veut réordonner les valeurs propres sur la diagonale de la matrice triangulaire supérieure T, issue de la décomposition de Schur de A, de façon à ce que des valeurs propres appartenant à un même bloc apparaissent à des places adjacentes sur la diagonale. Ceci peut se faire par transformations unitaires. En effet, le théorème de décomposition de Schur [12, p.17] stipule que la matrice Q peut être choisie de façon à ce que les valeurs propres apparaissent dans n'importe quel ordre pré-défini sur la diagonale de T. En fait, à cette étape, on ne fait que de recalculer une nouvelle factorisation de Schur, connaissant l'ordre d'apparition des valeurs propres sur la diagonale.

Principe du réordonnancement
Grâce au tri effectué à l'étape précédente, on peut déterminer une permutation σ de l'ensemble $E = \{1, 2, \ldots, n\}$ qui associe, à une valeur propre se trouvant à la $i^{\text{ième}}$ place sur la diagonale de la factorisation de Schur initiale, $\sigma(i)$ qui est sa place sur la diagonale de la nouvelle factorisation. Or, on sait que n'importe quelle permutation peut se décomposer en produits de transpositions agissant sur deux éléments adjacents. Dans le cas général, il suffit trouver la décomposition de la permutation en produit de transposition et d'utiliser le fait qu'une transposition sur deux éléments adjacents de la diagonale peut se faire par une transformation semblable unitaire.

Échange de deux éléments consécutifs sur la diagonale
Soit $T = \begin{pmatrix} \lambda_1 & t \\ 0 & \lambda_2 \end{pmatrix}$ avec $(\lambda_1, \lambda_2, t) \in \mathbb{C}^3$.
On recherche une matrice unitaire Q de la forme $\begin{pmatrix} a & -\bar{b} \\ b & a \end{pmatrix}$ vérifiant $Q^*TQ = \begin{pmatrix} \lambda_2 & t \\ 0 & \lambda_1 \end{pmatrix}$

- Q unitaire $\Longrightarrow \begin{cases} a\bar{a} + b\bar{b} = 1 \\ \bar{a}b - ab = 0 \end{cases}$
- $Q^* \begin{pmatrix} \lambda_1 & t \\ 0 & \lambda_2 \end{pmatrix} Q = \begin{pmatrix} \lambda_2 & t \\ 0 & \lambda_1 \end{pmatrix} \quad \Longrightarrow \quad a(\lambda_2 - \lambda_1) - bt = 0.$

On obtient comme solution

$$\begin{cases} a = \frac{\pm|t|}{\sqrt{|\lambda_2 - \lambda_1|^2 + |t|^2}} & b = a\frac{\lambda_2 - \lambda_1}{t} & \text{si } t \neq 0, \\ a = 0 & b = -1 & \text{si } t = 0. \end{cases}$$

Plus généralement, on sait donc déterminer une matrice Q unitaire telle que $T = Q^*AQ$ et telle que les valeurs propres sur la diagonale soient rangées dans l'ordre relatif aux différents blocs. C'est à dire que T peut s'écrire sous la forme

$$T = \begin{pmatrix} T_{11} & T_{12} & \ldots & T_{1q} \\ & T_{22} & \ldots & T_{2q} \\ & & \ddots & \vdots \\ & & & T_{qq} \end{pmatrix}$$

où les blocs diagonaux $T_{kk}, k = 1, \ldots, q$ satisfont alors aux deux conditions suivantes:

$$\forall \lambda_i \in \Lambda_0(T_{kk}) \;\; \exists \lambda_j \in \Lambda_0(T_{kk}) \;\; : \;\; |u_i^* u_j| \geq 1 - \eta \quad \text{pour } dim(T_{kk}) > 1 \tag{18}$$

$$\forall \lambda_i \in \Lambda_0(T_{kk}) \;\; \forall \lambda_j \notin \Lambda_0(T_{kk}) \;\; : \;\; |u_i^* u_j| < 1 - \eta \quad \text{pour } q > 1. \tag{19}$$

Application à l'exemple du paragraphe précédent
La détermination de la décomposition maximale admissible nous a permis de trouver un nombre de blocs $q = 2$ et la répartition des valeurs propres dans chacun des deux blocs suivants, $B_1 = \{\lambda_1, \lambda_3, \lambda_4\}$ et $B_2 = \{\lambda_2, \lambda_5\}$. De ces données, on déduit la permutation σ. Cette permutation σ se décompose en produit de deux transpositions τ_1 et τ_2 agissant sur deux éléments consécutifs.

$$\begin{array}{ll} 1 \xrightarrow{\sigma} 1 & 1 \xrightarrow{\tau_1} 1 \xrightarrow{\tau_2} 1 \\ 2 \longrightarrow 4 & 2 \longrightarrow 2 \longrightarrow 4 \\ 3 \longrightarrow 2 & 3 \longrightarrow 4 \longrightarrow 2 \\ 4 \longrightarrow 3 & 4 \longrightarrow 3 \longrightarrow 3 \\ 5 \longrightarrow 5 & 5 \longrightarrow 5 \longrightarrow 5 \\ & \xrightarrow[\sigma]{} \end{array}$$

D'où la transformation de la matrice T

$$T = \begin{pmatrix} \lambda_1 & \dots & \dots & \dots & \dots \\ & \lambda_2 & \dots & \dots & \dots \\ & & \lambda_3 & \dots & \dots \\ & & & \lambda_4 & \dots \\ & & & & \lambda_5 \end{pmatrix} \quad \Rightarrow \quad T = \begin{pmatrix} \boxed{\begin{matrix} \lambda_1 & \dots & \dots \\ & \lambda_3 & \dots \\ & & \lambda_4 \end{matrix}} & \begin{matrix} \dots & \dots \\ \dots & \dots \\ \dots & \dots \end{matrix} \\ T_{11} & \boxed{\begin{matrix} \lambda_2 & \dots \\ & \lambda_5 \end{matrix}} \\ & T_{22} \end{pmatrix}$$

Matrice T initiale — Matrice T issue du réordonnancement

Les valeurs propres de la matrice T issue du réordonnancement apparaissent bien dans l'ordre désiré, en première position sur la diagonale de T_{11} celles appartenant au premier bloc, ensuite sur la diagonale de T_{22} celles appartenant au second bloc, et ainsi de suite... On peut vérifier que les éléments propres de T_{11} et de T_{22} vérifient bien les conditions (18) et (19).

2.1.3 Élimination des blocs extra-diagonaux de la matrice T

Étude du cas $q = 2$ blocs

Théorème 4
Soit $T = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}$ *avec* $T_{11} \in \mathbb{C}^{m\times m}$, $T_{22} \in \mathbb{C}^{n\times n}$, $T_{12} \in \mathbb{C}^{m\times n}$. *Si* R *est solution de l'équation de Sylvester* $T_{11}R - RT_{22} = T_{12}$ *et* P *est de la forme* $P = \begin{pmatrix} I_m & -R \\ 0 & I_n \end{pmatrix}$, *alors*

$$P^{-1}TP = \begin{pmatrix} T_{11} & 0 \\ 0 & T_{22} \end{pmatrix}.$$

Nous voyons donc que l'élimination du terme extra-diagonal est liée à l'existence d'une solution de l'équation de Sylvester. Les deux théorèmes suivants vont nous donner des indications sur l'existence et l'unicité de telles solutions.

Théorème 5 (Existence)
Soit $A \in \mathbb{C}^{m\times m}$, $B \in \mathbb{C}^{n\times n}$, $C \in \mathbb{C}^{m\times n}$. L'équation de Sylvester $AX - XB = C$ admet une solution si et seulement si les matrices $\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ et $\begin{pmatrix} A & C \\ 0 & B \end{pmatrix}$ sont semblables.

La démonstration de ce théorème peut être trouvée dans [9, p.279].

Théorème 6 (Unicité)
Soit A,B et C définis comme dans le théorème 5.
Soit Φ linéaire : $\begin{array}{ccc} \mathbb{C}^{m\times n} & \to & \mathbb{C}^{m\times n} \\ X & \to & \Phi(X) = AX - XB \ . \end{array}$

$$\begin{array}{lcl} \textit{Alors } \Phi \textit{ non singulière} & \Longleftrightarrow & \Lambda_0(A) \cap \Lambda_0(B) = \emptyset \\ \textit{Or } \Phi \textit{ non singulière} & \Longleftrightarrow & \exists! X \textit{ solution de } AX - XB = C \ . \end{array}$$

Une démonstration de ce théorème dans le cas général peut être trouvée dans [12]. L'équation matricielle de Sylvester $AX - XB = C$ est étudiée en détail dans [4, p.228], [9, p.268].

Montrons que dans le cas où A et B sont les blocs diagonaux de la matrice T après réordonnancement (i.e. $A = T_{11}$ et $B = T_{22}$) et C le bloc extra-diagonal T_{12}, il existe nécessairement une solution à l'équation de Sylvester $AX - XB = C$, et donc l'élimination du terme extra-diagonal est toujours possible (voir théorème 4). En effet, deux cas sont alors possibles :

- Soit $\Lambda_0(A) \cap \Lambda_0(B) = \emptyset$, alors d'après le théorème 6, la solution non seulement existe, mais elle est unique.
- Soit $\exists \lambda \in \Lambda_0(A) \cap \Lambda_0(B)$. Ainsi, la valeur propre λ est présente dans deux blocs distincts. Notons (λ_1, u) le couple propre associé à la valeur propre λ du premier bloc et (λ_2, v) celui du second bloc. Les vecteurs u et v, normés à l'unité, sont des vecteurs propres associés à la valeur propre $\lambda = \lambda_1 = \lambda_2$. Du fait de l'appartenance de (λ_1, u) et de (λ_2, v) à des blocs distincts, d'après la propriété (19), on a $0 < |u^*v| < 1 - \eta$. Donc u, v forment une famille libre. Cela implique que la boite de Jordan,associée à la valeur propre λ, de la forme de Jordan de la matrice T, à la forme $\mathrm{diag}(\ldots, J_1(\lambda_1), J_2(\lambda_2), \ldots)$ où $J_1(\lambda_1)$ et $J_2(\lambda_2)$ sont des blocs de Jordan élémentaires associés à λ. Il est alors aisé de montrer que dans ces conditions les matrices $\tilde{T} = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ et $T = \begin{pmatrix} A & C \\ 0 & B \end{pmatrix}$ sont semblables, ce qui, d'après le théorème 5, implique qu'il existe une solution.
  En effet, soit J la forme de Jordan de la matrice T telle que :

$$J = \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix} = X^{-1} \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} X, \tag{20}$$

avec

- X de la forme $\begin{pmatrix} X_1 & X_2 \\ 0 & X_3 \end{pmatrix}$,
- les colonnes de X_1 étant formées des vecteurs propres et des vecteurs principaux de A,
- X_1 et X_3 inversibles,
- $J_1 = \mathrm{diag}(J_{11}, J_1(\lambda))$,
- $J_2 = \mathrm{diag}(J_2(\lambda), J_{22})$.

Le calcul par bloc de l'égalité (20) donne en particulier $J_1 = X_1^{-1}AX_1$ et $J_2 = X_3^{-1}BX_3$. On obtient finalement

$$\begin{aligned} T &= X \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix} X^{-1} \\ &= X \begin{pmatrix} X_1^{-1} & 0 \\ 0 & X_3^{-1} \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \begin{pmatrix} X_1 & 0 \\ 0 & X_3 \end{pmatrix} X^{-1} \\ &= Y\tilde{T}Y^{-1} \end{aligned}$$

avec $Y = \begin{pmatrix} I & X_2X_3^{-1} \\ 0 & I \end{pmatrix}$.

Ce qui revient à dire que T et $\tilde{T}$ sont semblables.

En conséquence, on peut affirmer que pour toute matrice T issue de notre procédure de réordonnancement, on peut effectuer l'élimination du terme extra-diagonal. Il est à noter que la matrice P du théorème 4 n'est pas une matrice unitaire.

Étude du cas où l'on a q blocs

Par application récursive des théorèmes 4, 6, 5, on peut en déduire un théorème de décomposition bloc-diagonale d'une matrice triangulaire supérieure par bloc.

Théorème 7 (Réduction sous forme diagonale)

Soit $T = \begin{pmatrix} T_{11} & T_{12} & \dots & T_{1q} \\ & T_{22} & \dots & T_{2q} \\ & & \ddots & \vdots \\ & & & T_{qq} \end{pmatrix}$ *une matrice triangulaire par bloc avec* T_{ii}, $i = 1, \dots, q$ *carrée. Si pour tout* $i < j$, *il existe* R *solution de l'équation de Sylvester* $T_{ii}R - RT_{jj} = -T_{ij}$, *alors il existe* $Y \in \mathbb{C}^{n\times n}$ *non singulière telle que*

$$Y^{-1}TY = diag(T_{11}, T_{22}, \dots, T_{qq})$$

Preuve: Soit à annuler le terme en position $(i,j) \quad i<j$.

Soit $Y_{ij} = \begin{pmatrix} & & & j & \\ I_1 & & & \vdots & \\ & I_2 & & \boxed{Z_{ij}} & \cdots \\ & & \ddots & & \\ & & & \ddots & \\ & & & & I_q \end{pmatrix} \quad i$.

Posons $\bar{T} = [\bar{T}_{ij}] = Y_{ij}^{-1} T Y_{ij} \quad i<j$.
Alors T et $\bar{T}$ ne diffèrent que par les termes :

$$\begin{cases} \bar{T}_{iw} = T_{iw} - Z_{ij}T_{jw} & w = j+1,\ldots,q\ , \\ \bar{T}_{kj} = T_{kj} + T_{ki}Z_{ij} & k = 1,\ldots,i-1\ , \\ \bar{T}_{ij} = T_{ij} - Z_{ij}T_{jj} + T_{ii}Z_{ij}\ . \end{cases}$$

Notre but est d'annuler le coefficient $\bar{T}_{ij}$, on obtient alors

$$T_{ii}Z_{ij} - Z_{ij}T_{jj} = -T_{ij}, \tag{21}$$

équation de Sylvester d'inconnue Z_{ij} qui admet une solution au moins d'après les hypothèses du théorème 7. Pour éliminer le bloc en position (i,j), il suffit de déterminer Z_{ij} en résolvant l'équation (21), d'annuler effectivement le bloc en position (i,j) et de remettre à jour les blocs des positions $(i,w) \quad w = j+1,\ldots,q$ et $(k,j) \quad k = 1,\ldots,i-1$ (cf. figure 2).

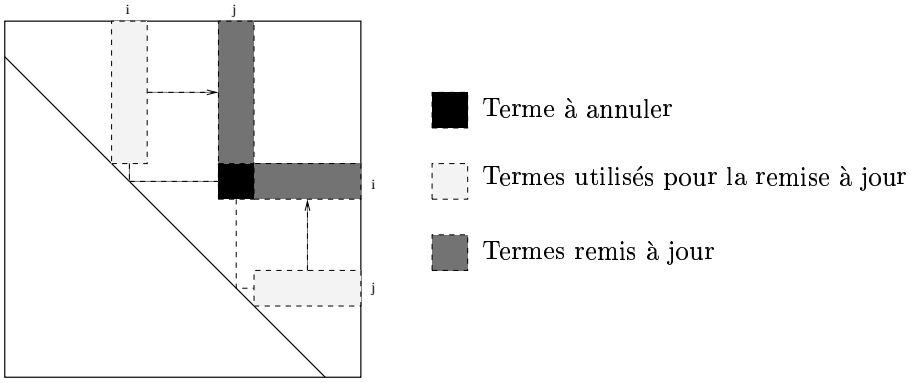

FIG. 2: *Relations des dépendences pour l'élimination de l'élément* (i,j)

On fait cela pour chacun des blocs extra-diagonaux. Il faut cependant procéder avec méthodologie, car l'ordre d'élimination est important, il faut annuler les termes $T_{ij} \quad i<j$ en prenant soin de ne pas retoucher à des termes que l'on a déjà annulés. Pour cela on

choisit d'éliminer les blocs ligne par ligne par ordre croissant (i.e. on commence par ceux de la première ligne, puis ceux de la seconde et ainsi de suite...). Pour une ligne donnée, on élimine les blocs colonne par colonne par ordre croissant (i.e. pour une ligne i donnée, on va éliminer dans l'ordre, le bloc en position $(i, i+1)$, puis celui en position $(i, i+2)$ et ainsi de suite jusqu'au dernier bloc de la ligne). C'est l'ordre indiqué sur la figure 3. De cette façon, lorsqu'un bloc en position (i, j) est annulé, on n'y retouche plus ultérieurement (cf. figure 4).

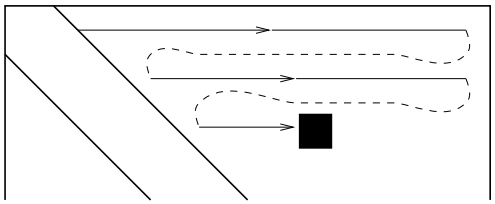

FIG. 3: *Ordre d'élimination des éléments extra-diagonaux*

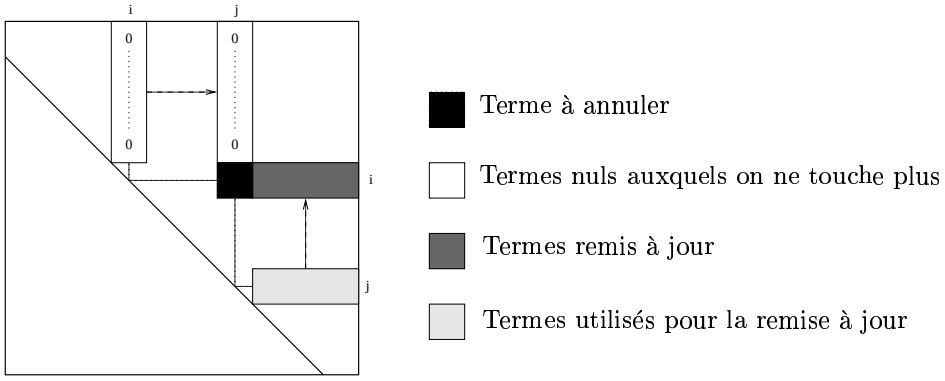

FIG. 4: *Élimination des termes extra-diagonaux de façon ordonnée*

On a donc ainsi démontré le théorème 7. À la fin de cette étape, on a obtenu la factorisation par bloc de la matrice A sous la forme :

$$A = S\mathrm{diag}(T_{11}, T_{22}, \ldots, T_{qq})S^{-1}$$

□

2.1.4 Orthonormalisation des blocs-colonnes de S

On a vu l'importance que l'on portait à ce que le conditionnement de la matrice S ne soit pas trop grand. Dans cette optique, et pour que $K(S)$ ne soit pas trop éloigné de $K(S_{OPT})$, (voir inégalité (11) du théorème 2), on va orthonormaliser chacun des blocs-colonnes S_i. Ainsi une nouvelle bloc-diagonalisation sera obtenue. Pour cela on va considérer la décomposition QR de chaque S_i,

$$S_1 = Q_1 R_1 \quad \dots \quad S_q = Q_q R_q \quad \forall i = 1, \dots, q \quad Q_i^* Q_i = Id_i \tag{22}$$

et posons $Q = [Q_1|\dots|Q_q]$. Grâce à la factorisation (22), on a :

$$\begin{aligned}
A &= S\mathrm{diag}(T_{11}, T_{22}, \dots, T_{qq})S^{-1} \\
&= [S_1|\dots|S_q]\mathrm{diag}(T_{11}, T_{22}, \dots, T_{qq})[S_1|\dots|S_q]^{-1} \\
&= [Q_1R_1|\dots|Q_qR_q]\mathrm{diag}(T_{11}, T_{22}, \dots, T_{qq})[Q_1R_1|\dots|Q_qR_q]^{-1} \\
&= [Q_1|\dots|Q_q]\mathrm{diag}(R_1, \dots, R_q)\mathrm{diag}(T_{11}, T_{22}, \dots, T_{qq})[\mathrm{diag}(R_1, \dots, R_q)]^{-1}[Q_1|\dots|Q_q]^{-1} \\
&= [Q_1|\dots|Q_q]\mathrm{diag}(R_1T_{11}R_1^{-1}, \dots, R_qT_{qq}R_q^{-1})[Q_1|\dots|Q_q]^{-1}
\end{aligned}$$

Pour obtenir la décomposition finale (23) de A, il suffit de remplacer S par Q et $\mathrm{diag}(T_{11}, T_{22}, \dots, T_{qq})$ par $\mathrm{diag}(R_1T_{11}R_1^{-1}, \dots, R_qT_{qq}R_q^{-1})$. Si on note D_i les blocs diagonaux finaux, on a la décomposition finale

$$A = S\mathrm{diag}(D_1, \dots, D_q)S^{-1} \tag{23}$$

2.1.5 Réduction itérative du nombre de bloc

Notre technique de bloc-diagonalisation est basée sur le concept suivant : on choisit les blocs de façon à ce que lors de l'élimination du bloc extra-diagonal (i, j), la norme $\|Z_{ij}\|$ de la matrice d'élimination solution de (21) ne soit pas trop grande. Cependant, le nombre de blocs étant variable, il se peut que les multiples éliminations engendrent, malgré nos précautions, un conditionnement de la matrice S qui soit jugé comme étant trop grand. Une première méthode consiste alors à recommencer le processus de bloc-diagonalisation, mais avec un $\eta' > \eta$, d'où une réduction du nombre de blocs et ainsi un conditionnent moins grand. Cependant, nous avons que vu la détermination du choix des blocs n'est valable que pour $\eta \ll 1$, ce qui n'est plus nécéssairement le cas pour η'. De ce fait, nous avons opté pour une seconde méthode qui consiste à réduire itérativement le nombre de blocs, ceci jusqu'à ce que la valeur de $\kappa(S)$ soit considérée comme raisonnable. Pour cela on va déterminer deux blocs que l'on va fusionner en un seul.

Exemple de réduction de $q = 3$ à $q = 2$ blocs
Considérons les trois blocs:

- $S_1 = [\alpha_1|\dots|\alpha_{i_1}] \quad S^1 = \mathrm{eng}(\alpha_i, i = 1, \dots, i_1)$

- $S_2 = [\beta_1|\dots|\beta_{i_2}] \quad S^2 = \mathrm{eng}(\beta_i, i = 1, \dots, i_2)$
- $S_3 = [\gamma_1|\dots|\gamma_{i_3}] \quad S^3 = \mathrm{eng}(\gamma_i, i = 1, \dots, i_3)$

avec $S_j^* S_j = I_j \quad j = 1, 2, 3 \quad$ (i.e. S_j base orthonormale de S^j)
Posons $\mathcal{V}_{ij} = \mathcal{V}(S^i, S^j)$.
Alors $\cos(\mathcal{V}_{ij}) = \|S_i^* S_j\| \quad \left\{ \begin{array}{l} \text{i=1,2,3} \\ \text{j=1,2,3} \end{array} \right.$
On a trois possibilités de réduction possible :

$$\boxed{S_1} \qquad \boxed{S_2|S_3} \qquad \cos(\mathcal{V}(\tilde{S}_1, \tilde{S}_2)) \geq \max(\cos(\mathcal{V}_{12}), \cos(\mathcal{V}_{13})) \tag{24}$$

$$\boxed{S_1|S_2} \qquad \boxed{S_3} \qquad \cos(\mathcal{V}(\tilde{S}_1, \tilde{S}_2)) \geq \max(\cos(\mathcal{V}_{13}), \cos(\mathcal{V}_{23})) \tag{25}$$

$$\underbrace{\boxed{S_1|S_3}}_{\tilde{S}_1} \qquad \underbrace{\boxed{S_2}}_{\tilde{S}_2} \qquad \cos(\mathcal{V}(\tilde{S}_1, \tilde{S}_2)) \geq \max(\cos(\mathcal{V}_{12}), \cos(\mathcal{V}_{23})) \tag{26}$$

Nous allons étudier en détail le cas de la première fusion, les deux autres donnant des calculs identiques. Fusionnons dans un unique bloc, les blocs numéro deux et trois. Nous avons donc :

- $\tilde{S}_1 = S_1$ base orthonormale de S_1
- $\tilde{S}_2$ base orthonormale du sous-espace vectoriel engendré par S^2 et S^3.

Soit $S = [\tilde{S}_1|\tilde{S}_2]$, on a $\tilde{S}_i^* \tilde{S}_i = I_i, \quad i = 1, 2$. Explicitons l'angle entre les sous-espaces $\tilde{S}^1$ et $\tilde{S}^2$, engendrés respectivement par les colonnes de $\tilde{S}_1$ et de $\tilde{S}_2$:

$$\begin{aligned} \tilde{\mathcal{V}}_{12} &= \mathcal{V}(\tilde{S}_1, \tilde{S}_2) \\ &= \min\{arccos|u^*v|/u \in \tilde{S}_1, v \in \tilde{S}_2, \|u\| = \|v\| = 1\} \\ &\leq \min(\min\{arccos|u^*v|/u \in S_1, v \in S_2, \|u\| = \|v\| = 1\}, \\ &\qquad \min\{arccos|u^*v|/u \in S_1, v \in S_3, \|u\| = \|v\| = 1\}) \\ &\leq \min(\mathcal{V}_{12}, \mathcal{V}_{13}) \ . \end{aligned}$$

En prenant le cosinus de chacun des deux membres de l'inégalité précédente, on en déduit l'inégalité (24):

$$\cos(\min(\mathcal{V}_{12}, \mathcal{V}_{13})) = \max(\cos(\mathcal{V}_{12}), \cos(\mathcal{V}_{13})) \leq \cos(\mathcal{V}(\tilde{S}_1, \tilde{S}_2)) = \cos(\tilde{\mathcal{V}}_{12})$$

Les inégalités (25) et (26) se démontrent exactement de la même façon.
On a maintenant $q = 2$ blocs. Alors grâce à l'égalité (14) du théorème 3, on connait la valeur du conditionemment de la matrice S,

$$K(S) = \cot(\frac{\mathcal{V}(\tilde{S}_1, \tilde{S}_2)}{2}) \ . \tag{27}$$

Notre but étant d'obtenir un conditionnement le plus petit possible, on choisira la décomposition qui donne la plus petite valeur pour $\cot(\frac{\mathcal{V}(\tilde{S}_1,\tilde{S}_2)}{2})$, ou ce qui revient au même, pour $\cos(\mathcal{V}(\tilde{S}_1,\tilde{S}_2))$. Cependant, le calcul explicite, pour chacune des fusions possibles, des bases orthonormales des sous-espaces $\tilde{S}^1$ et $\tilde{S}^2$ est très couteux en terme de calcul. Bien que réalisable dans le cas où l'on passe de trois à deux blocs, on décide d'éviter ces calculs.
On se contente de minimiser la borne inférieure $\alpha_i,\ i = 1,2,3$ issue des trois inégalités (24),(25),(26) de $\cos(\mathcal{V}(\tilde{S}_1,\tilde{S}_2))$. Par exemple, supposons que $\cos(\mathcal{V}_{13}) \leq \cos(\mathcal{V}_{12}) \leq \cos(\mathcal{V}_{23})$, les trois inégalités deviennent:

$$\cos(\mathcal{V}(\tilde{S}_1,\tilde{S}_2)) \geq \max(\cos(\mathcal{V}_{12}),\cos(\mathcal{V}_{13})) = \cos(\mathcal{V}_{12}) = \alpha_1$$

$$\cos(\mathcal{V}(\tilde{S}_1,\tilde{S}_2)) \geq \max(\cos(\mathcal{V}_{13}),\cos(\mathcal{V}_{23})) = \cos(\mathcal{V}_{23}) = \alpha_2$$

$$\cos(\mathcal{V}(\tilde{S}_1,\tilde{S}_2)) \geq \max(\cos(\mathcal{V}_{12}),\cos(\mathcal{V}_{23})) = \cos(\mathcal{V}_{23}) = \alpha_3$$

Comme on a $\alpha_1 \leq \alpha_2 = \alpha_3$, la fusion qui minimisera la borne inférieure de $\cos(\mathcal{V}(\tilde{S}_1,\tilde{S}_2))$ dans cet exemple est celle où l'on fusionne les blocs initialement numérotés deux et trois. Plus généralement, lors de la fusion de trois à deux blocs, ce minimum est atteint lorsque l'on fusionne les deux blocs dont le cosinus de l'angle entre les sous-espaces qu'ils représentent est le plus grand. Ce critère de choix des blocs à fusionner est très peu coûteux en terme de calcul, cependant la fusion qu'il engendre n'est pas nécéssairement optimale. En effet, on ne minimise que $\alpha_i,\ i = 1,2,3$, borne inférieure des $\cos(\mathcal{V}(\tilde{S}_1,\tilde{S}_2))$ sur l'ensemble des fusions possibles. Rien ne nous garantit que ce choix sera effectivement celui qui minimisera $\cos(\mathcal{V}(\tilde{S}_1,\tilde{S}_2))$ sur l'ensemble des fusions possibles. Rendre cette étape optimale est possible, mais coûteut, il faut alors calculer pour chacune des trois fusions possibles une base orthonormale de $\tilde{S}^1$ et $\tilde{S}^2$, calculer ensuite $\cos(\mathcal{V}(\tilde{S}_1,\tilde{S}_2)) = \|\tilde{S}_1^*\tilde{S}_2\|$, et choisir la fusion qui minimise $\cos(\mathcal{V}(\tilde{S}_1,\tilde{S}_2))$, ce qui minimisera (27).

Étude de la réduction de q à $q-1$ blocs
L'idée directrice est la même que dans le paragraphe précédent. Soit $T = \mathrm{diag}(T_{11},\ldots,T_{qq})$, $S = [S_1|\ldots|S_q]$ avec $S_j^*S_j = I_j \quad j = 1,\ldots,q$. On réduit le nombre de blocs en fusionnant dans le bloc i, les blocs i et j avec $i < j$ (voir figure 5). Une fois la fusion effectuée, on se retrouve dans la situation suivante :
soit $f(k) = \begin{cases} k & k = 1,\ldots,i-1,i+1,\ldots,j-1 \\ k+1 & k = j,\ldots,q-1\ , \end{cases}$
et $\begin{cases} \tilde{S}_k = S_{f(k)} \quad k = 1,\ldots,i-1,i+1,\ldots,q-1 \\ \tilde{S}_i = \text{base orthonormale de } \mathrm{eng}(S^i,S^j)\ , \end{cases}$
alors la nouvelle structure par bloc de S est $S = [\tilde{S}_1|\ldots|\tilde{S}_{q-1}]$ avec $\tilde{S}_j^*\tilde{S}_j = I_j,\ j = 1,\ldots,q-1$.

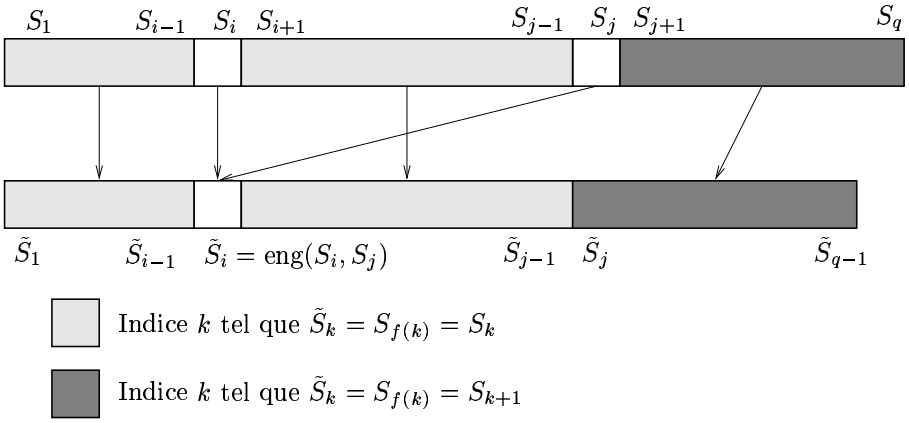

FIG. 5: *Structure par bloc de la matrice S lors de la fusion de q à $q-1$ blocs*

De l'inégalité (12) du théorème 2, on peut déduire l'inégalité suivante :

$$\max_k(csc(\tilde{\mathcal{V}}_k)) \leq K(S) \leq q \max_k(csc(\tilde{\mathcal{V}}_k)) \ . \tag{28}$$

De plus, d'après la définition de $\tilde{\mathcal{V}}_k$ on a $\tilde{\mathcal{V}}_k = \mathcal{V}(\tilde{S}^k, \mathrm{eng}_{w\neq k}\{\tilde{S}^w\}) \leq \min_{w\neq k}(\tilde{\mathcal{V}}_{kw})$

$\Rightarrow \quad csc(\tilde{\mathcal{V}}_k) \geq csc(\min_{w\neq k}(\tilde{\mathcal{V}}_{kw}) = \max_{w\neq k}(csc(\tilde{\mathcal{V}}_{kw}))$

$\Rightarrow \quad \max_k(csc(\tilde{\mathcal{V}}_k)) \geq \max_{w\neq k}(csc(\tilde{\mathcal{V}}_{kw}))$.

Or pour

$$\begin{cases} w \neq k \\ k \neq i \\ w \neq i \end{cases} \quad csc(\tilde{\mathcal{V}}_{wk}) = csc(\mathcal{V}_{f(w)f(k)}) \tag{29}$$

$$k \neq i \quad csc(\tilde{\mathcal{V}}_{ik}) \geq \max(csc(\mathcal{V}_{if(k)}), csc(\mathcal{V}_{jf(k)})) \ . \tag{30}$$

On a donc en utilisant l'égalité (29) et l'inégalité (30)

$$\begin{aligned} \max_{k\neq w} csc(\tilde{\mathcal{V}}_{kw}) &= \max\{ \max_{\substack{w\neq k \\ k\neq i \\ w\neq i}} csc(\tilde{\mathcal{V}}_{wk}), \max_{k\neq i} csc(\tilde{\mathcal{V}}_{ik})\} \\ &\geq \max\{ \max_{\substack{w\neq k \\ k\neq i \\ w\neq i}} csc(\mathcal{V}_{f(w)f(k)}), \max_{k\neq i} (csc(\mathcal{V}_{if(k)}), csc(\mathcal{V}_{jf(k)}))\} \ . \end{aligned}$$

Posons : $\alpha = \max\{ \max_{\substack{w \neq k \\ k \neq i \\ w \neq i}} csc(\mathcal{V}_{f(w)f(k)}), \max_{k \neq i}(csc(\mathcal{V}_{if(k)}), csc(\mathcal{V}_{jf(k)}))\}$.

On obtient les inégalités suivantes :

$$\max_{k}(csc(\tilde{\mathcal{V}}_k)) \geq \max_{k \neq w}(csc(\tilde{\mathcal{V}}_{kw})) \geq \alpha \ . \tag{31}$$

Or pour minimiser $K(S)$, il faut d'après l'inégalité (28) minimiser le terme $\max_{k}(csc(\tilde{\mathcal{V}}_k))$. Pour des raisons de coût de calcul, on se borne à minimiser α, borne inférieure de cette quantité. C'est dans le choix des blocs i et j que l'on décide de fusionner, que va résider l'éventuelle minimisation de α. Pour $w \neq i$ on a $f(w) \in \{1, \ldots, i-1\} \cup \{i+1, \ldots, j-1\} \cup \{j+1, \ldots, q\}$. Soit $(\lambda, \mu) \quad \lambda < \mu$ définis par $csc(\mathcal{V}_{\lambda\mu}) = \max_{i \neq j}(csc(\mathcal{V}_{ij}))$. Étudions quel choix des indices i et j va rendre α minimum.

- Premier choix possible: $(i, j) = (\lambda, \mu)$

$$\alpha = \max\{ \max_{\substack{w \neq k \\ k \neq \lambda \\ w \neq \lambda}} csc(\underbrace{\mathcal{V}_{f(w)f(k)}}_{\neq \mathcal{V}_{\lambda\mu}}), \max_{k \neq \lambda}(csc(\underbrace{\mathcal{V}_{\lambda f(k)}}_{\neq \mathcal{V}_{\lambda\mu}}), csc(\underbrace{\mathcal{V}_{\mu f(k)}}_{\neq \mathcal{V}_{\lambda\mu}}))\}$$

  or comme $csc(\mathcal{V}_{\lambda\mu}) = \max_{k \neq w}(csc(\mathcal{V}_{kw}))$ on a

$$\alpha \leq csc(\mathcal{V}_{\lambda\mu}) \ . \tag{32}$$

- Second choix possible: $(i, j) \neq (\lambda, \mu)$

$$\alpha = \max\{ \max_{\substack{w \neq k \\ k \neq i \\ w \neq i}} csc(\mathcal{V}_{f(w)f(k)}), \max_{k \neq i} csc(\mathcal{V}_{if(k)}, \mathcal{V}_{jf(k)})\}$$

  Étudions alors les différentes possibilités:

  Si $i \neq \lambda$ et $j \neq \mu \quad \Rightarrow \max_{\substack{w \neq k \\ k \neq i \\ w \neq i}} csc(\mathcal{V}_{f(w)f(k)}) = csc(\mathcal{V}_{\lambda\mu})$

  Si $i = \lambda$ et $j \neq \mu \quad \Rightarrow \max_{k \neq \lambda} csc(\mathcal{V}_{\lambda f(k)}) = csc(\mathcal{V}_{\lambda\mu})$

  Si $i \neq \lambda$ et $j = \mu \quad \Rightarrow \max_{k \neq i} csc(\mathcal{V}_{\mu f(k)}) = csc(\mathcal{V}_{\lambda\mu})$

  Dans tout les cas on obtient le résultat suivant:

$$\alpha = csc(\mathcal{V}_{\lambda\mu}) \ . \tag{33}$$

Finalement, si on veut minimiser α, au vue des résultats (32) et (33), il faut faire le choix de fusionner les blocs d'indices i et j tel que $(i,j) = (\lambda,\mu)$. On en déduit un critère de détermination des indices des blocs à fusionner : *on décide de fusionner les deux blocs dont le csc de l'angle des sous-espaces qu'ils représentent est le plus grand.* De la même façon que lors de la réduction de trois à deux blocs, cette étape n'est pas optimale, en ce sens qu'on ne fait que minimiser α borne inférieure de la quantité $\max_k(csc(\tilde{\mathcal{V}}_k))$ que l'on cherche à minimiser. Cependant cette façon de faire est très peu coûteuse et donne, comme on le verra lors de l'analyse des résultats numériques, de bonnes décompositions.
Rendre cette étape optimale est très difficile, voire impossible lorsque le nombre de blocs q est élevé. En effet, il faudrait alors pour chacune des $\frac{q(q-1)}{2}$ fusions possibles calculer $\max_k(csc(\tilde{\mathcal{V}}_k))$. Ceci implique la détermination d'une base orthonormale de $\tilde{S}_i$, $i = 1,\dots q-1$, ainsi que de son complémentaire, pour chacune des fusions possibles...

Cette réduction itérative du nombre de blocs est répétée jusqu'à ce que l'on obtienne un conditionnement de la matrice S que l'on considère comme raisonnable.

2.2 Algorithme de bloc-diagonalisation de A

On se donne $A \in \mathbb{C}^{n\times n}$, $\epsilon \in \mathbb{R}^{+*}$, $K_{limite} > 1$: Conditionnement maximal admissible.

1. **Factorisation de Schur**: $A = QTQ^*$
   - Q unitaire
   - T triangulaire supérieure : $T = \begin{pmatrix} \lambda_1 & \dots & \dots & \dots \\ & \lambda_2 & \dots & \dots \\ & & \ddots & \dots \\ & & & \lambda_n \end{pmatrix}$
2. **Calcul des vecteurs propres de A**
   $$U = [u_1|u_2|\dots|u_n] \text{ tels que } \forall i = 1,\dots,n \begin{cases} Au_i = \lambda_i u_i \\ \|u_i\| = 1 \end{cases}$$
3. **Construction du graphe** G sur $E = \{1,2,\dots,n\}$ défini par:
   $$(i,j) \in G \iff |u_i^* u_j| \geq 1 - \eta$$
4. **Détermination des différentes composantes connexes de G**
   $$\Rightarrow \begin{cases} q = \text{ nombre de blocs } = \text{nombre de composantes connexes différentes} \\ \text{Répartition des différents } \lambda_i \text{ dans chaque bloc} \end{cases}$$
5. **Réordonnancement des valeurs propres sur la diagonale**
   Calcul de la nouvelle décomposition de Schur ordonnée selon les blocs $A = QTQ^*$
   - Q unitaire

– T triangulaire supérieure : $T = \begin{pmatrix} T_{11} & \dots & \dots & \dots \\ & T_{22} & \dots & \dots \\ & & \ddots & \dots \\ & & & T_{qq} \end{pmatrix}$

6. **Élimination des termes extra-diagonaux de T**

   – $A = SDS^{-1}$

   – $S = [S_1|\dots|S_q]$

   – D diagonale par bloc $= D = \begin{pmatrix} T_{11} & & & \\ & T_{22} & & \\ & & \ddots & \\ & & & T_{qq} \end{pmatrix}$

7. **Orthonormalisation de chaque S_i** $\quad i = 1, \dots, q$

   – $A = SDS^{-1}$

   – $S = [S_1|\dots|S_q] \quad S_i^* S_i = Id_i \quad \forall i = 1, \dots, q$

   – D diagonale par bloc $= D = \begin{pmatrix} D_1 & & & \\ & D_2 & 0 & \\ & & \ddots & \\ & 0 & & D_q \end{pmatrix}$

8. **Calcul et test de $K(S)$**

   – Si $K(S) \leq K_{limite}$ alors
   $D = S^{-1}AD$ est la décomposition désirée $\Rightarrow$ **FIN**

   – Si $K(S) > K_{limite}$ alors
   Réduction du nombre de blocs ($q \rightarrow q-1$)

   - Recherche des indices μ et λ vérifiant $csc(\mathcal{V}_{\lambda\mu}) = \max_{i \neq j} csc(\mathcal{V}_{ij})$
   - Fusion des blocs numéro λ et μ

   $\Rightarrow \quad \begin{cases} q := q-1 \text{ nouveau nombre de blocs} \\ \text{Nouvelle répartition des différents } \lambda_i \text{ dans chaque bloc} \end{cases}$

   Retour au point 5

2.3 Résultats numériques

2.3.1 Présentation des matrices exemples

– **Matrice de Frank**

- Cette matrice est tirée de [8]. C'est une matrice Hessenberg supérieure, dont les valeurs propres sont très mal conditionnées. L'ordre de cette matrice dans nos exemples sera de 50. Elle a la structure suivante :

$$M = \begin{bmatrix} \ddots & \dots & \dots & 4 & 3 & 2 & 1 \\ \ddots & \ddots & & \vdots & \vdots & \vdots & \vdots \\ & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ & & 4 & 4 & \vdots & \vdots & \vdots \\ & & & 3 & 3 & \vdots & \vdots \\ & & & & 2 & 2 & \vdots \\ & & & & & 1 & 1 \end{bmatrix}$$

- Soit V la matrice des vecteurs propres. On a alors $K(V) = 5.55e + 10$

– **Matrice GRCAR**

- Cette matrice est tirée de [8]. C'est une matrice de Toeplitz, dont les valeurs propres sont sensibles aux perturbations. L'ordre de cette matrice dans nos exemples sera de 50. Elle a la structure suivante :

$$M = \begin{bmatrix} 1 & 1 & 1 & 1 & & & \\ -1 & 1 & 1 & 1 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & 1 & 1 & 1 & 1 \\ & & & -1 & 1 & 1 & 1 \\ & & & & -1 & 1 & 1 \\ & & & & & -1 & 1 \end{bmatrix}$$

- Soit V la matrice des vecteurs propres. On a alors $K(V) = 2.01e + 8$

– **Matrice Pentoep**

- Cette matrice est tirée de [8]. C'est une matrice de Toeplitz, dont les valeurs propres sont sensibles aux perturbations. L'ordre de cette matrice dans nos exemples sera de 50. Elle a la structure suivante :

$$M = \begin{bmatrix} 0 & 0 & 1 & & & \\ 0.5 & 0 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & 1 \\ & & & \ddots & \ddots & 0 \\ & & & & 0.5 & 0 \end{bmatrix}$$

- Soit V la matrice des vecteurs propres. On a alors $K(V) = 4.11e+11$

– **Matrice Aléatoire triangulaire supérieure**

- C'est une matrice aléatoire, triangulaire supérieure, à coefficients complexes.
- Soit V la matrice des vecteurs propres. On a alors $K(V) = 5.39e+11$

– **Matrice de Frank par bloc**

- Soit B la matrice diagonale par bloc, dont chaque bloc est une matrice de Frank d'ordre 10 shiftée d'une certaine valeur. Soit Q une matrice unitaire quelconque. On construit la matrice A de la façon suivante : $A = Q^*BQ$.
- Soit V la matrice des vecteurs propres. On a alors $K(V) = 3.45e+06$

2.3.2 Résultats numériques de bloc-diagonalisation

– La méthode 1 correspond à un choix de blocs où le critère de sélection est l'angle entre les vecteurs propres.

– La méthode 2 est celle de Stewart [1].

– La méthode 3 est celle qui correspond à un choix de blocs basé sur la distance entre les valeurs propres.

Lorsque que le résultat n'a pas pu être calculé par une méthode, la valeur de $K(S)$ dans le tableau des résultats est remplacée par /.

- Résultats numériques de la bloc-diagonalisation de la matrice GRCAR(50)

Nombres de blocs	$K(S)$ avec tri méthode 1	$K(S)$ avec tri méthode 2	$K(S)$ avec tri méthode 3
13	8825	/	7.69e+05
12	7036	4.756e+04	2.588e+05
11	3903	/	2.469e+05
10	3409	3.186e+04	2.335e+05
9	2167	2.307e+04	2.206e+05
8	1976	/	2.039e+05
7	1684	/	1.815e+05
6	1469	/	3.16e+04
5	1372	/	2.677e+04
4	1359	/	2.169e+04
3	702.8	/	1.51e+04
2	308.5	9.522e+04	8070
1	1	1	1

TAB. 1: *Comparaison de l'efficacité des différents tris sur la matrice* $GRCAR(50)$

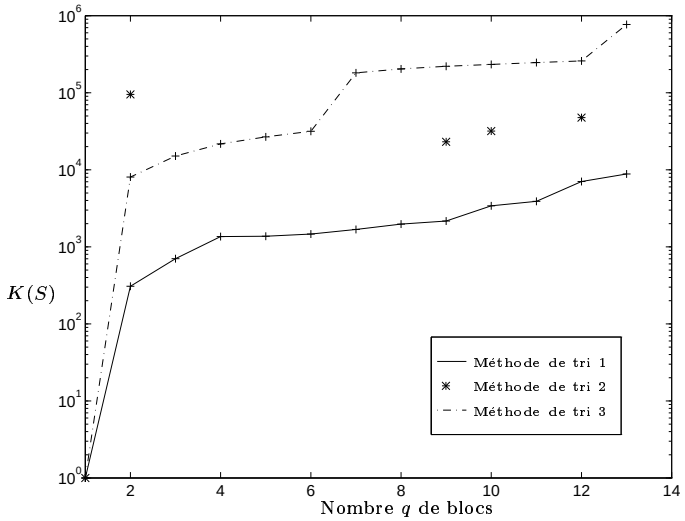

FIG. 6: *Évolution du conditionnement* $K(S)$ *en fonction du nombre* q *de blocs pour les trois tris étudiés, sur la matrice* $GRCAR(50)$

- Résultats numériques de la bloc-diagonalisation de la matrice FRANK(50)

Nombres de blocs	$K(S)$ avec tri méthode 1	$K(S)$ avec tri méthode 2	$K(S)$ avec tri méthode 3
9	9.876e+05	6.463e+07	2.197e+06
8	2.096e+04	/	2.083e+06
7	2615	4.871e+05	1.876e+06
6	2626	1.272e+05	1.494e+06
5	667.8	/	9.437e+05
4	199.2	3150	4.164e+05
3	186.8	1201	1.077e+05
2	47.81	/	1.13e+04
1	1	1	1

TAB. 2: *Comparaison de l'éfficacité des différents tris sur la matrice* $FRANK(50)$

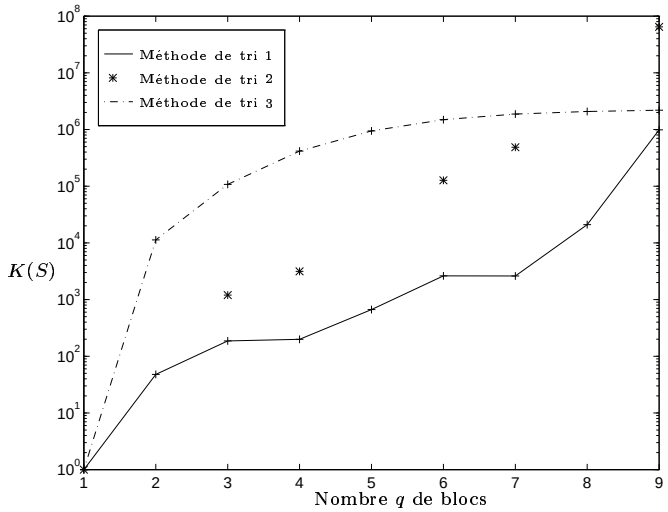

FIG. 7: *Évolution du conditionnement* $K(S)$ *en fonction du nombre* q *de blocs pour les trois tris étudiés, sur la matrice* $FRANK(50)$

- Résultats numériques de la bloc-diagonalisation de la matrice Pentoep(50)

Nombres de blocs	$K(S)$ avec tri méthode 1	$K(S)$ avec tri méthode 2	$K(S)$ avec tri méthode 3
10	4.559e+05	/	4.559e+05
9	4.375e+05	/	4.375e+05
8	3.586e+05	/	3.586e+05
7	1.165e+05	/	1.165e+05
6	1.168e+05	/	1.168e+05
5	9.655e+04	/	9.655e+04
4	2.562e+04	/	2.562e+04
3	2.158e+04	/	2.158e+04
2	1.301e+04	7.17e+08	1.301e+04
1	1	1	1

TAB. 3: *Comparaison de l'éfficacité des différents tris sur la matrice Pentoep*(50)

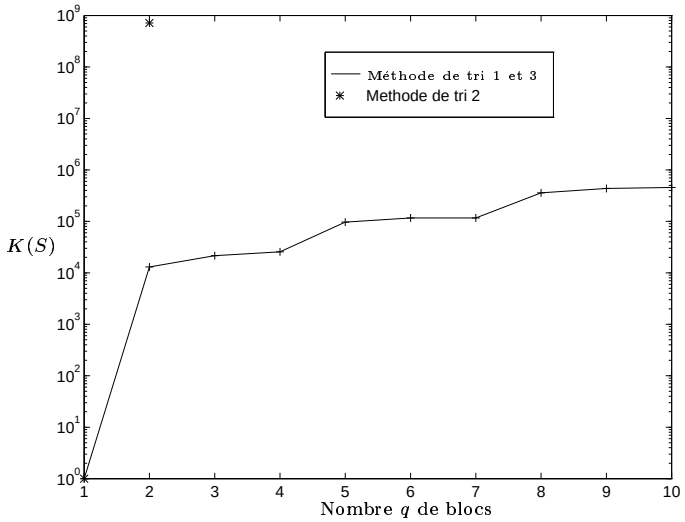

FIG. 8: *Évolution du conditionnement $K(S)$ en fonction du nombre q de blocs pour les trois tris étudiés, sur la matrice $PENTOEP(50)$*

- Résultats numériques de la bloc-diagonalisation de la matrice Aléatoire(50)

Nombres de blocs	$K(S)$ avec tri méthode 1	$K(S)$ avec tri méthode 3
10	4.158e+04	1.341e+09
9	2.339e+04	1.319e+09
8	1.252e+04	1.24e+09
7	9523	4.727e+05
6	3927	4.387e+05
5	2634	6570
4	452.3	5913
3	342.5	2065
2	47.53	47.53
1	1	1

TAB. 4: *Comparaison de l'efficacité des différents tris sur une matrice aléatoire triangulaire supérieure à coefficients complexes*

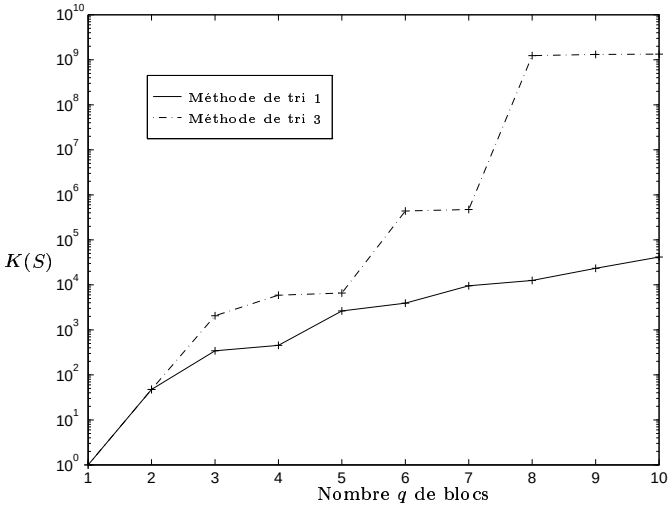

FIG. 9: *Évolution du conditionnement $K(S)$ en fonction du nombre q de blocs pour les trois tris étudiés, sur la matrice $ALEATOIRE(50)$*

- Résultats numériques de la bloc-diagonalisation de la matrice de Frank par bloc

Nombres de blocs	$K(S)$ avec tri méthode 1	$K(S)$ avec tri méthode 3
10	23.39	24.52
9	15.86	24.28
8	15.59	24.24
7	14.38	17.32
6	8.907	15.48
5	7.55	15.49
4	7.345	10.6
3	3.59	6.72
2	2.667	3.419
1	1	1

TAB. 5: *Comparaison de l'efficacité des différents tris sur la matrice de Frank par bloc*

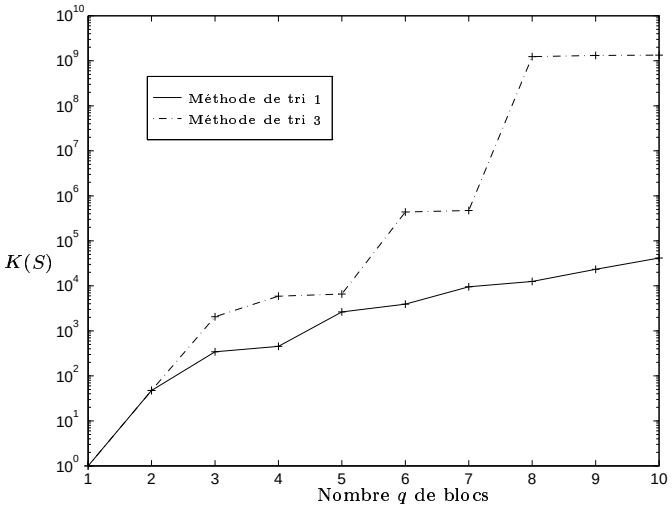

FIG. 10: *Évolution du conditionnement $K(S)$ en fonction du nombre q de blocs pour les trois tris étudiés, sur la matrice Frank par bloc*

2.3.3 Commentaires et conclusions

Le but de notre bloc-diagonalisation est de calculer une décomposition du type (8) avec un conditionnement $\kappa(S)$ le plus petit possible. À un nombre de blocs q fixé, la décomposition qui sera considérée comme la meilleure sera celle dont le conditionnement $\kappa(S)$ sera le plus proche de 1. Les résultats de l'évolution de $\kappa(S)$ en fonction du nombre q de blocs, pour les trois méthodes de tri considérées sont présentés de la figure 6 à la figure 10.
On peut tout d'abord remarquer que la méthode 2 n'est adaptée qu'à des matrices à coefficients reéls. C'est pourquoi elle n'apparaît pas dans la figure 9 où la matrice est à coefficients complexes. De plus, de part la façon même dont est effectuée la bloc-diagonalisation avec cette méthode, il n'est pas facile, voire parfois impossible d'obtenir une factorisation, pour un nombre q de blocs donné. Ainsi le symbole "/" dans les entrées des tableaux résultats représente ces échecs.
L'analyse des résultats révèle que pour un nombre de blocs q donné, le conditionnemnt $\kappa(S)$ issu de la factorisation utilisant la méthode de tri 1 est toujours inférieur ou égal à ceux issus des factorisations basées sur les méthodes de tri 2 ou 3, et ce d'un facteur variant de 1 à 1000. Dans le cas particulier de la matrice de PENTOEP, les résultats sont identiques que l'on utilise la méthode de tri 1 ou 3. Par contre, il s'avère souvent impossible de trouver une bloc-diagonalisation avec un conditionnement $\kappa(S)$ raisonnable si on utilise la méthode de tri 2.
Lorsque la factorisation basée sur la méthode de tri 2 donne un résultat, celui-ci est souvent meilleur qu'une factorisation issue du tri 3, mais moins bon que celle issue du tri 1.
En conclusion, il semble que notre procédure (tri 1) de bloc-diagonalisation est relativement performante, au sens où le conditionnement $\kappa(S)$ obtenu est bien inférieur à celui trouvé par d'autres procédures (tri 2 ou tri 3), à un nombre de blocs q fixé. Bien que non optimale (voir paragraphe 2.1.5), les résultats peuvent être considérés comme satisfaisants. De plus, la flexibilité de notre algorithme permet, une fois la décomposition maximale en q_{max} blocs calculée, de déduire pour un coût modique n'importe quelle décomposition en q blocs avec $q < q_{max}$, grace à notre méthode de fusion des blocs.

3 Approximation du portrait spectral de A

Dans cette section, nous supposons connaître déjà une factorisation bloc-diagonale de type (8) de la matrice A, avec un conditionnement $K(S)$ inférieur à une valeur seuil fixée par l'utilisateur. Montrons comment l'utiliser pour calculer une approximation du portrait spectral de A.

3.1 Approximation utilisant le portrait spectral de D

Considérons la décomposition (8) de la matrice A qui peut s'écrire $A = SDS^{-1}$. Alors $A - zI = S(D - zI)S^{-1}$ d'où

$$\begin{aligned}
\sigma_{min}(A - zI) &= \sigma_{min}(S(D - zI)S^{-1}) \\
&= \frac{1}{\sigma_{max}(S(D - zI)^{-1}S^{-1})} \\
&= \frac{1}{\|S(D - zI)^{-1}S^{-1}\|} \\
&\geq \frac{1}{\|S\|\|(D - zI)^{-1}\|\|S^{-1}\|} \\
&= \frac{1}{K(S)\ \sigma_{max}((D - zI)^{-1})} \\
&= \frac{1}{K(S)}\sigma_{min}(D - zI)\ .
\end{aligned}$$

De manière identique, on a $D - zI = S^{-1}(A - zI)S$, d'où

$$\begin{aligned}
\sigma_{min}(D - zI) &= \sigma_{min}(S^{-1}(A - zI)S) \\
&\geq \frac{1}{K(S)}\ \sigma_{min}(A - zI)\ .
\end{aligned}$$

Finalement on obtient l'encadrement

$$\frac{1}{K(S)}\sigma_{min}(D - zI) \leq \sigma_{min}(A - zI) \leq K(S)\sigma_{min}(D - zI)\ . \tag{34}$$

C'est cet encadrement qui est à l'origine de la proposition suivante :

Proposition 1 *Soit* $\epsilon \geq 0$, $\tilde{\epsilon}_1 = \frac{\|A\|}{\kappa(S)\|D\|}\epsilon$ *et* $\tilde{\epsilon}_2 = \frac{\kappa(S)\|A\|}{\|D\|}\epsilon$ *alors*

$$\Lambda_{\tilde{\epsilon}_1}(D) \subset \Lambda_\epsilon(A) \subset \Lambda_{\tilde{\epsilon}_2}(D)\ . \tag{35}$$

Soit $\epsilon_i = \frac{\|D\|}{\|D_i\|}\epsilon$ *alors*

$$\Lambda_\epsilon(D) = \cup_{i=1}^{q}\Lambda_{\epsilon_i}(D_i)\ . \tag{36}$$

Preuve: Rappelons les définitions des différents ϵ-pseudospectres qui interviennent dans la proposition 1:

- $\Lambda_\epsilon(A) = \{z \in \mathbb{C} : \sigma_{min}(A - zI) \leq \epsilon\|A\|\}$
- $\Lambda_{\tilde{\epsilon}_1}(D) = \{z \in \mathbb{C} : \sigma_{min}(D - zI) \leq \tilde{\epsilon}_1\|D\|\}$
- $\Lambda_{\tilde{\epsilon}_2}(D) = \{z \in \mathbb{C} : \sigma_{min}(D - zI) \leq \tilde{\epsilon}_2\|D\|\}$

- Soit $z \in \Lambda_{\tilde{\epsilon}_1}(D)$, alors

$$\sigma_{min}(D - zI) \leq \frac{\|A\|}{K(S)\|D\|}\epsilon\|D\| \ . \tag{37}$$

D'après l'encadrement (34) et en utilisant l'inégalité (37), on obtient :

$$\begin{aligned} \sigma_{min}(A - zI) &\leq K(S)\sigma_{min}(D - zI) \\ &\leq \epsilon\|A\| \ , \end{aligned}$$

donc $z \in \Lambda_\epsilon(A)$ et par conséquent $\Lambda_{\tilde{\epsilon}_1}(D) \subset \Lambda_\epsilon(A)$.

- De manière identique, soit $z \in \Lambda_\epsilon(A)$.
Alors

$$\sigma_{min}(A - zI) \leq \epsilon\|A\| \tag{38}$$

D'après l'encadrement (34) et en utilisant l'inégalité (38), on obtient :

$$\begin{aligned} \sigma_{min}(D - zI) &\leq K(S)\sigma_{min}(A - zI) \\ &\leq \frac{\kappa(S)\|A\|}{\|D\|}\epsilon\|D\| \ , \end{aligned}$$

donc $z \in \Lambda_{\tilde{\epsilon}_2}(D)$ et par conséquent $\Lambda_\epsilon(A) \subset \Lambda_{\tilde{\epsilon}_2}(D)$.
D'où l'inclusion (35) de la proposition 1.

Notons qu'en utilisant l'encadrement

$$\frac{\|D\|}{K(S)} \leq \|A\| \leq K(S)\|D\| \ , \tag{39}$$

on peut déduire de (35) l'inclusion suivante

$$\Lambda_{\tilde{\epsilon}_1}(D) \subset \Lambda_\epsilon(A) \subset \Lambda_{\tilde{\epsilon}_2}(D) \tag{40}$$

où $\epsilon \geq 0$, $\tilde{\epsilon}_1 = \frac{\epsilon}{(\kappa(S))^2}$ et $\tilde{\epsilon}_2 = \epsilon(\kappa(S))^2$. Cependant, cette inclusion (40) est beaucoup moins précise que (35).

- Démontrons maintenant l'égalité (36)
  Comme $D = \mathrm{diag}(D_1, D_2, \ldots, D_q)$ on a

$$\sigma_{min}(D) = \min_i(\sigma_{min}(D_i)) \ . \tag{41}$$

Soit $z \in \Lambda_\epsilon(D)$, cela implique

$$\sigma_{min}(D - zI) \leq \epsilon \, \|D\| \ . \tag{42}$$

Or comme $D - zI$ est diagonale par bloc, d'après l'égalité (41), on a

$$\sigma_{min}(D - zI) = \min_i(\sigma_{min}(D_i - zI_i)) \ . \tag{43}$$

En utilisant l'inégalité (42), on en déduit

$$\min_i(\sigma_{min}(D_i - zI_i)) \leq \epsilon \, \|D\| \ . \tag{44}$$

De l'inégalité (44), on peut affirmer
$\exists j \in \{1, \ldots, q\}$ tel que

$$\begin{aligned} \sigma_{min}(D_j - zI_j) &\leq \epsilon \, \|D\| \\ &\leq \frac{\|D\|}{\|D_j\|} \epsilon \|D_j\| \end{aligned}$$

$$\Rightarrow \quad z \in \Lambda_{\epsilon_j}(D_j) \quad \Rightarrow \quad z \in \cup_{i=1}^q \Lambda_{\epsilon_i}(D_i)$$

On a donc montré que

$$\Lambda_\epsilon(D) \subset \cup_{i=1}^q \Lambda_{\epsilon_i}(D_i) \ . \tag{45}$$

Réciproquement, soit $z \in \cup_{i=1}^q \Lambda_{\epsilon_i}(D_i)$, alors $\exists j \in \{1, \ldots, q\}$ tel que $z \in \Lambda_{\epsilon_j}(D_j)$, ce qui implique

$$\sigma_{min}(D_j - zI_j) \leq \epsilon_j \, \|D_j\| = \epsilon \, \|D\| \ . \tag{46}$$

Donc en utilisant successivement l'égalité (43) et l'inégalité (46), on obtient

$$\begin{aligned} \sigma_{min}(D - zI) &= \min_i(\sigma_{min}(D_i - zI_i)) \\ &\leq \sigma_{min}(D_j - zI_j) \\ &\leq \epsilon \, \|D\| \ . \end{aligned}$$

Donc $z \in \Lambda_\epsilon(D)$ et on a donc montré que

$$\cup_{i=1}^q \Lambda_{\epsilon_i}(D_i) \subset \Lambda_\epsilon(D) \ . \tag{47}$$

Des inclusion (45) et (47), on en déduit l'égalité (36) de la proposition 1, ce qui termine la démonstration de cette proposition. □

Explicitons les résultats de la proposition 1. L'inclusion (35) de cette proposition nous montre que $\Lambda_\epsilon(A)$ et donc le portrait spectral de la matrice A peut être approché en calculant $\Lambda_{\tilde{\epsilon}_1}(D)$ et $\Lambda_{\tilde{\epsilon}_2}(D)$. Or ces ensembles sont beaucoup plus facile à calculer, ceci grâce à l'égalité (36). En effet, le ϵ-pseudospectre et donc par suite le portrait spectral de la matrice D peut être obtenu à partir de chacun des $\Lambda_{\epsilon_i}(D_i)$, $i = 1, \ldots, q$. Comme la taille des matrices D_i, $i = 1, \ldots, q$ est petite, des méthodes fiables basées sur la décomposition en valeurs singulières peuvent être utilisées pour calculer leur ϵ-pseudospectres. De plus, comme le calcul des $\Lambda_\epsilon(D_i)$, $i = 1, \ldots, q$ est totalement indépendant les uns des autres, celui-ci peut se faire en parallèle.
Cependant, l'inclusion (35) étant très large, si on décide d'approcher $\Lambda_\epsilon(A)$ par $\Lambda_{\tilde{\epsilon}}(D)$, alors la question de la valeur de $\tilde{\epsilon}$ se pose inévitablement. En effet l'inclusion (35) nous indique seulement que l'on doit prendre $\tilde{\epsilon}_1 \leq \tilde{\epsilon} \leq \tilde{\epsilon}_2$, ce qui dans la grande majorité des cas est un encadrement très large. Nous visualiserons ces faits, dans le paragraphe 4.4, sur différents exemples numériques.

3.2 Approximation utilisant le champ des valeurs de D

Un autre concept pour étudier les propriétés spectrales d'une matrice est l'utilisation du champ des valeurs [9]. Le champ des valeurs d'une matrice A est défini par

$$F(A) = \{u^* A u : \ u \in \mathbb{C}^n, \ \|u\| = 1\} . \tag{48}$$

Cet ensemble est un fermé borné de $\mathbb{C}$ (voir [9, p.8]). Il possède en plus la propriété d'être convexe et contient l'enveloppe convexe du spectre de la matrice A (voir [9, p.17]). Les relations reliant le champ des valeur et le ϵ-pseudospectre de A font l'objet de la proposition suivante.

Proposition 2 *Pour $\epsilon \geq 0$ on a*

$$\Lambda_\epsilon(A) \subset F(D) + \Delta_{\epsilon K(S)\|A\|} \tag{49}$$

$$\Lambda_\epsilon(A) \subset F(A) + \Delta_{\epsilon\|A\|} \tag{50}$$

où $\Delta_\tau = \{z \in \mathbb{C} : \ |z| \leq \tau\}$, et

$$F(D) = Co\,(\cup_{i=1}^q F(D_i)) \tag{51}$$

où $Co\,(\cup_{i=1}^q F(D_i))$ représente l'enveloppe convexe de $\cup_{i=1}^q F(D_i)$.

Preuve: Soit $z \in \Lambda_\epsilon(A)$, par définition du ϵ-spectre :

$$\exists u, \Delta \quad \|u\| = 1 \quad \text{tels que} \quad (A + \Delta)u = zu \quad \text{avec} \quad \|\Delta\| \leq \epsilon\|A\| ,$$

soit encore

$$z = \underbrace{u^*Au}_{\in F(A)} + \underbrace{u^*\Delta u}_{z_1} \ .$$

On en déduit $|z_1| = |u^*\Delta u| \leq \|\Delta\| \leq \epsilon\|A\|$, donc $z_1 \in \Delta_{\epsilon\|A\|}$, ce qui démontre (50).
De la même façon, soit $z \in \Lambda_\epsilon(A)$, on a :

$$\exists u, \Delta \quad \|u\| = 1 \quad \text{tels que} \quad (A+\Delta)u = zu \quad \text{avec} \quad \|\Delta\| \leq \epsilon\|A\| \ ,$$

ce qui peut se mettre sous la forme $zu = (SDS^{-1} + \Delta)u$, ou encore :

$$(D + S^{-1}\Delta S)S^{-1}u = zS^{-1}u \ .$$

Posons $v = S^{-1}u$, on obtient :

$$v^*(D + S^{-1}\Delta S)v = zv^*v \ ,$$

soit encore

$$z = \underbrace{\frac{v^*Dv}{v^*v}}_{\in F(D)} + \underbrace{\frac{v^*S^{-1}\Delta Sv}{v^*v}}_{z_1} \ .$$

Le calcul de la norme de z_1 donne :

$$|z_1| = \frac{|v^*S^{-1}\Delta Sv|}{|v^*v|} \leq \frac{\|v^*\|\|S^{-1}\Delta S\|\|v\|}{\|v\|^2} \leq K(S)\|\Delta\| \leq K(S)\epsilon\|A\| \ ,$$

donc $z_1 \in \Delta_{\epsilon K(S)\|A\|}$, ce qui démontre (49).
Pour la démonstration de (51) se reporter à [9, p.12]. □

De la même façon qu'à la proposition 1, (51) signifie que le champ des valeurs de la matrice D peut facilement être déterminé à partir de celui des matrices (D_i), $i = 1, \ldots, q$.

3.3 Quelques propriétés intéressantes

Quand on approche le portrait spectral de A par celui de la matrice bloc-diagonale D, on peut se demander pour quel ϵ les portraits spectraux de deux blocs différents D_i et D_j se superposent ? La réponse est donnée par l'opérateur sep_λ définit pour deux matrices M_1 et M_2 dans [14, 3] par

$$\text{sep}_\lambda(M_1, M_2) = \inf_\lambda \max(\sigma_{min}(M_1 - \lambda I),\ \sigma_{min}(M_2 - \lambda I)). \tag{52}$$

En d'autres termes, soit E et F des perturbations respectives de M_1 et de M_2. Alors $\text{sep}_\lambda(M_1, M_2) = \max(\|E\|, \|F\|)$ telles que E et F soient les plus petites perturbations

vérifiant $\Lambda_0(M_1+E)\cap\Lambda_0(M_2+F)\neq\emptyset$.
Il est clair que

$$\mathrm{sep}_\lambda(M_1,M_2)\leq dist\left(\Lambda_0(M_1),\Lambda_0(M_2)\right)/2\ , \tag{53}$$

où $dist\left(\Lambda_0(M_1),\Lambda_0(M_2)\right)$ est la distance entre le spectre de M_1 et celui de M_2. Notons que l'on a pour $z\in\mathbb{C}$ et $i=1,2$

$$\begin{aligned} dist(F(M_i),z) &= \inf_{\|u\|=1}|u^*M_iu-z| \\ &= \inf_{\|u\|=1}|u^*(M_i-zI)u| \\ &\leq \inf_{\|u\|=1}\|(M_i-zI)u\| \\ &\leq \sigma_{min}(M_i-zI)\ . \end{aligned} \tag{54}$$

Ainsi, d'après (54) on a

$$\begin{aligned} dist(F(M_1),F(M_2)) &\leq dist(F(M_1),z)+dist(F(M_2),z) \quad \forall z\in C\subset\mathbb{C} \\ &\leq \sigma_{min}(M_1-zI)+\sigma_{min}(M_2-zI) \quad \forall z\in C\ , \end{aligned} \tag{55}$$

c'est donc vrai en particulier pour la borne inférieure de (55),

$$\begin{aligned} dist(F(M_1),F(M_2)) &\leq \inf_\lambda\left[\sigma_{min}(M_1-\lambda I)+\sigma_{min}(M_2-\lambda I)\right] \\ &\leq 2\inf_\lambda\left[\max(\sigma_{min}(M_1-\lambda I),\sigma_{min}(M_2-\lambda I))\right] \\ &\leq 2\mathrm{sep}_\lambda(M_1,M_2)\ . \end{aligned} \tag{56}$$

Finalement de (56) et (53), on obtient la proposition suivante

Proposition 3 *$\forall i,j=1,\ldots,q$ on a*

$$dist(F(D_i),F(D_j))/2\leq sep_\lambda(D_i,D_j)\leq dist\left(\Lambda_0(D_i),\Lambda_0(D_j)\right)/2\ . \tag{57}$$

4 Résultats numériques et complexité

4.1 Calcul effectif du portrait spectral d'une matrice A

Comme nous l'avons déjà vu, le portrait spectral de A consiste en la représentation des ensembles $\Lambda_\epsilon(A)=\{z\in\mathbb{C}:\sigma_{min}(A-zI)\leq\epsilon\|A\|\}$ pour plusieurs valeurs différentes de ϵ . En fait, pour des raisons de commodité, c'est la fonction $sp_A(z)=\log_{10}\left(\frac{\sigma_{min}(zI-A)}{\|A\|}\right)$ que l'on va visualiser. Pour cela, on se donne $[x_{min},x_{max},y_{min},y_{max}]$, ce qui va définir une zone du plan complexe, limite de la partie intéressante du portrait spectral. Ce domaine va être discrétisé en une grille de $N\times N$ points, représentée par les points z_{kw} d'affixe

$z_{kw} = x_{min} + k * p_x + i(y_{min} + w * p_y)$ avec $p_x = \left(\frac{x_{max}-x_{min}}{N-1}\right)$ et $p_y = \left(\frac{y_{max}-y_{min}}{N-1}\right)$. Soit M la matrice de terme générique $\log_{10}\left(\frac{\sigma_{min}(A-z_{kw}I)}{\|A\|}\right)$ (i.e. c'est en fait la transformation des points de la grille par la fonction $sp_A(z)$). On visualise le portrait spectral de A par les courbes de niveaux (d'élévation égale à $\log_{10}(\epsilon)$)des coefficients de la matrice M (cf. figure 11). Chaque courbe de niveau correspond à une valeur différente de ϵ. On prend généralement ϵ variant de 10^{-1} à 10^{-10} (i.e. $\log_{10}(\epsilon)$ variant de -1 à -10).

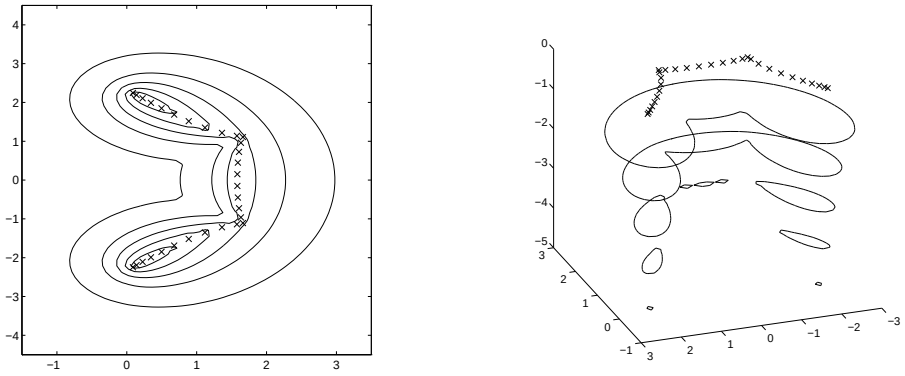

FIG. 11: *Exemples de portraits spectraux en 2D et en 3D*

4.2 Cas particulier d'une matrice diagonale par bloc

La seule différence avec le paragraphe précédent est qu'ici la matrice D est diagonale par bloc (i.e. $D = \mathrm{diag}(D_1, D_2, \ldots, D_q)$). On va se servir de cette propriété pour réduire les coûts de calcul.
La matrice $D - z_{kw}I$ étant une matrice diagonale par bloc, on peut utiliser l'égalité (41) pour $z = z_{kw}$, on obtient

$$\sigma_{min}(D - z_{kw}I) = \min_i(\sigma_{min}(D_i - z_{kw}I_i)) \ . \tag{58}$$

Le portrait spectral de D est la représentation, par l'intermédiaire de courbes de niveaux, des coefficiants de la matrice $M = \left[\frac{\sigma_{min}(D-z_{kw}I)}{\|D\|}\right]$. Cette dernière peut se calculer en utilisant l'opérateur $\min(,)$ définit pour des matrices de la façon suivante:

$$\begin{aligned} \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n} &\longrightarrow \mathbb{C}^{n\times n} \\ A, B &\longrightarrow C = \min(A, B) \end{aligned}$$

si $A = [a_{kw}]$, $B = [b_{kw}]$ alors $C = [c_{kw}]$ est definit par $c_{kw} = \min(a_{kw}, b_{kw})$.
L'opérateur $\min(,)$ ainsi défini, la matrice M peut se calculer en fonction des matrices

$M_i = \left[\frac{\sigma_{min}(D_i - z_{kw} I_i)}{\|D\|}\right]$ en vertu de l'égalité:

$$M = \min_{i=1,\cdots,q} M_i \ . \tag{59}$$

Dans le cas standard, le coût de calcul de la matrice M est de l'ordre de $O(N^2 n^3)$ (voir paragraphe 4.3). Dans le cas où la matrice est diagonale par bloc, en supposant que les blocs sont de même taille (i.e. de l'ordre de $\frac{N^2}{q}$), le coût de calcul de M est réduit à $O\left(\frac{N^2 n^3}{q^2}\right)$ (cf. paragraphe 4.3).

4.3 Complexité

Comme nous l'avons souligné au début du rapport, notre but est d'approcher le portrait spectral de A, par une méthode moins coûteuse en terme de calcul que dans le cas classique. Comparons donc la complexité de notre approche avec celle de la méthode standard.
Rappelons que le calcul de $sp_A(z)$ pour une matrice A d'ordre n, par la méthode de décomposition en valeurs singulières, est de l'ordre de $O(n^3)$, ceci pour chaque z d'une grille du plan complexe. Supposons que cette grille soit discrétisée en $N \times N$ points, alors le coût de calcul de $\Lambda_\epsilon(A)$ sur cette grille est de l'ordre de $O(N^2 n^3)$. Maintenant, supposons que nous avons effectué une bloc-diagonalisation de la matrice A, et ainsi obtenu q blocs D_i, $i = 1, \ldots, q$, chacun de taille n/q. Alors le coût séquentiel de notre approche est de l'ordre de $O(N^2 q(\frac{n}{q})^3)$, plus le coût d'une bloc-diagonalisation de A. De plus, si on utilise le fait que le calcul de chaque $\Lambda_\epsilon(D_i)$, $i = 1, \ldots, q$, peut être fait en parallèle, alors le facteur $O(N^2 q(\frac{n}{q})^3)$ du coût séquentiel devient $O(N^2(\frac{n}{q})^3)$. On peut encore faire mieux, en effet, au lieu de calculer $\Lambda_\epsilon(D_i)$ sur la totalité de la grille, on peut réduire le domaine de calcul à des sous-grilles. Théoriquement, en supposant que le domaine discrétisé en $N \times N$ points puisse être divisé en q^2 sous-domaines ne se chevauchant pas, de taille $\frac{N}{q} \times \frac{N}{q}$ points, alors le coût précédent est réduit à $O((\frac{N}{q})^2(\frac{n}{q})^3)$.

4.4 Résultats numériques

Visualisation de l'inclusion (35) de la proposition 1
La relation (35) de la proposition 1 nous donne un encadrement de $\Lambda_\epsilon(A)$ par l'intermédiaire de $\Lambda_{\tilde{\epsilon}_1}(D)$ et $\Lambda_{\tilde{\epsilon}_2}(D)$. Visualisons donc sur différents exemples numériques cette inclusion.

- Prenons tout d'abord pour A la matrice de Frank par bloc d'ordre 50. On éffectue plusieurs bloc-décompositions de la matrice A du type $A = SD_qS^{-1}$, où la matrice D_q est respectivement composée de $q = 3, 7, 12, 13, 14$ blocs. C'est ce que l'on visualise sur les figures (12), (13) et (14) pour $q = 7, 12, 14$.

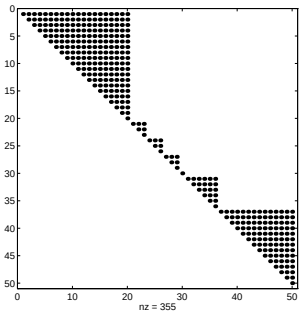

FIG. 12: *q=7*

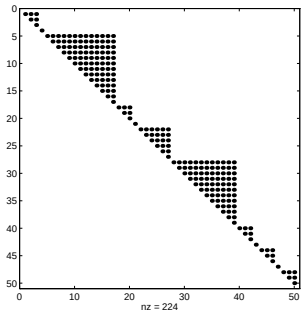

FIG. 13: *q=12*

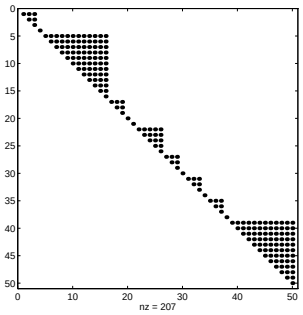

FIG. 14: *q=14*

Les figures suivantes (15),(16),(17),(18),(19),(20) représentent l'inclusion (35). $\Lambda_\epsilon(A)$ y est représenté en trait plein, pour une valeur fixée de $\epsilon = 10^{-2.5}$. Les autres courbes en pointillés sont les représentations respectives de $\Lambda_{\tilde{\epsilon}_1}(D_q)$ et $\Lambda_{\tilde{\epsilon}_2}(D_q)$. Dans cet exemple, on a fixé ϵ et on a visualisé l'inclusion (35) pour différentes bloc-diagonalisations de la matrice A. Il est intérressant de noter que lorsque le nombre de blocs de D est élevé (i.e. lorsque le conditionnement $K(S)$ est trés grand (>5000)), l'inclusion est trés peu précise et n'a pas grand intérêt. Par contre, dès que le nombre de blocs de D engendre un conditionnement raisonnable, cette inclusion devient beaucoup plus précise et peut être utilisée comme étant une première approximation grossière du portrait spectral de la matrice A.

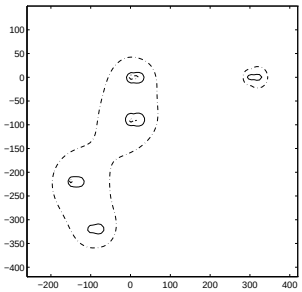

FIG. 15: *q=4*

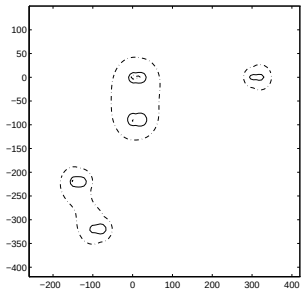

FIG. 16: *q=7*

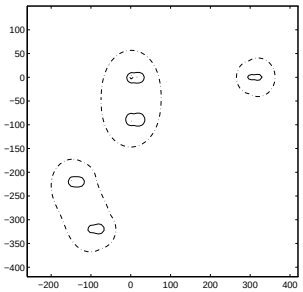

FIG. 17: *q=10*

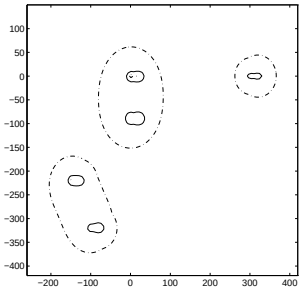

FIG. 18: *q=12*

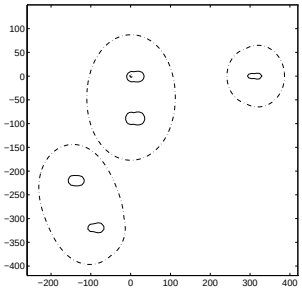

FIG. 19: *q=13*

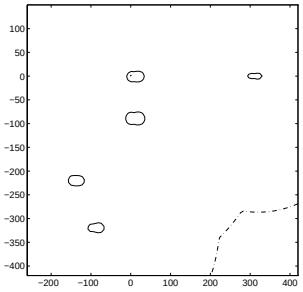

FIG. 20: *q=14*

- Même exemple, mais en prenant cette fois pour A la matrice de GRCAR, bloc-diagonalisée en $q = 3, 6$ et 13 blocs. On a pris dans cet exemple $\epsilon = 10^{-4.5}$.

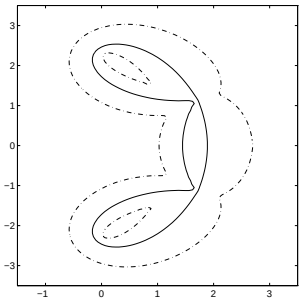

FIG. 21: *q=3*

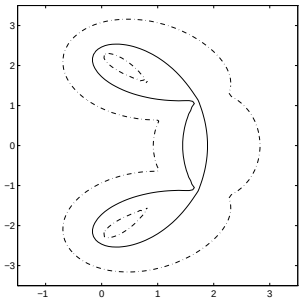

FIG. 22: *q=6*

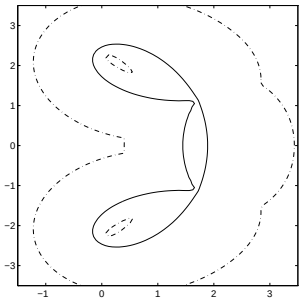

FIG. 23: *q=13*

- Dans un deuxième temps, on va toujours visualiser l'inclusion (35), en faisant varier ϵ, la décomposition de la matrice A en une matrice diagonale par bloc étant cette fois fixée. Plus précisément, soit A la matrice de GRCAR d'ordre 50. On fait une bloc-décomposition de A du type (8), où D est formée de 6 blocs diagonaux. On visualise alors sur chacune des figures (24),...,(29) les ensembles $\Lambda_\epsilon(A)$ en trait plein ainsi que $\Lambda_{\tilde{\epsilon}_1}(D)$ et $\Lambda_{\tilde{\epsilon}_2}(D)$ en pointillés, pour $\epsilon = 10^{-3.5}, \ldots, 10^{-8.5}$.

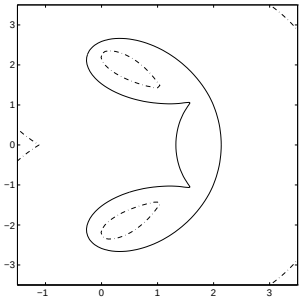

FIG. 24: $\epsilon = 10^{-3.5}$

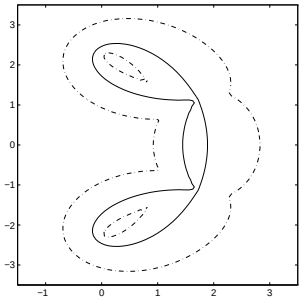

FIG. 25: $\epsilon = 10^{-4.5}$

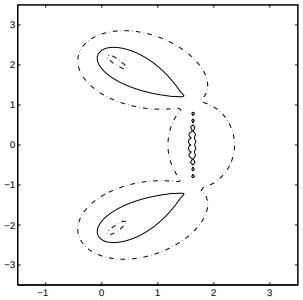

FIG. 26: $\epsilon = 10^{-5.5}$

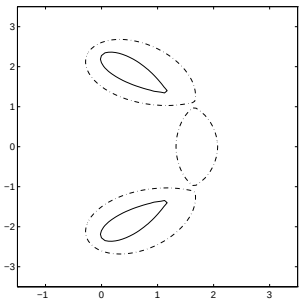

FIG. 27: $\epsilon = 10^{-6.5}$

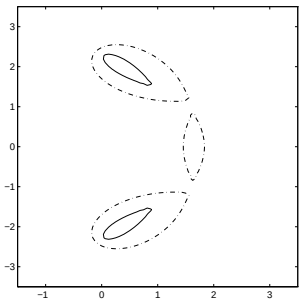

FIG. 28: $\epsilon = 10^{-7.5}$

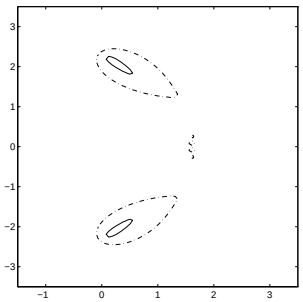

FIG. 29: $\epsilon = 10^{-8.5}$

- Enfin, dans ce dernier exemple concernant l'inclusion (35), nous allons visualiser, toujours pour une bloc-diagonalisation de A en 6 blocs, l'évolution de $\Lambda_{\tilde{\epsilon}}(D)$ pour $\tilde{\epsilon_1} \leq \tilde{\epsilon} \leq \tilde{\epsilon_2}$. Posons $\tilde{\epsilon} = \alpha\epsilon$. On visualise $\Lambda_\epsilon(A)$ dans la partie inférieure de chacune des figures. Dans la partie supérieure, on visualise $\Lambda_{\tilde{\epsilon}}(D)$.

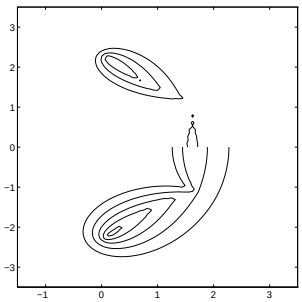

FIG. 30: $\alpha = 10^{-3.2}$

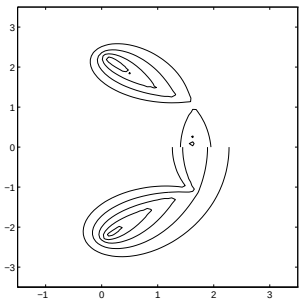

FIG. 31: $\alpha = 10^{-2.1}$

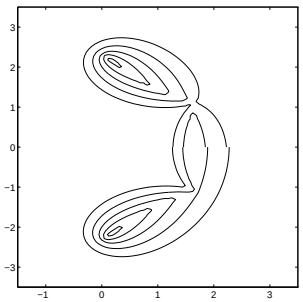

FIG. 32: $\alpha = 10^{-1.1}$

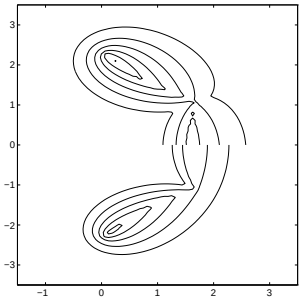

FIG. 33: $\alpha = 10^{2.33.10^{-4}}$

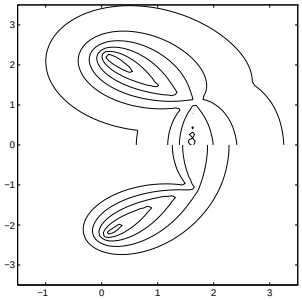

FIG. 34: $\alpha = 10^{1.1}$

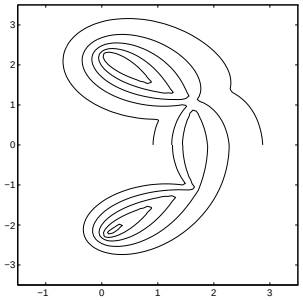

FIG. 35: $\alpha = 10^{2.1}$

- Si on se place dans l'optique d'approximer $\Lambda_\epsilon(A)$, l'exemple précédent, figures (30),...,(35), montre que parmi toutes les valeurs de $\tilde{\epsilon}$ vérifiant

$$\frac{\|A\|}{\kappa(S)\|D\|} \quad \leq \quad \tilde{\epsilon} \quad \leq \quad \frac{\kappa(S)\|A\|}{\|D\|} \ , \tag{60}$$

certaines valeurs de $\tilde{\epsilon}$ induisent une meilleure approximation de $\Lambda_\epsilon(A)$ par $\Lambda_{\tilde{\epsilon}}(D)$ que d'autres (i.e. figure(33)).

Cependant, le problème consistant à trouver la valeur $\tilde{\epsilon}_{Opt}$ pour laquelle l'approximation est la meilleure semble bien difficile à résoudre. Pour éffectuer cette étude, nous avons besoin de matérialiser la notion de «$\Lambda_{\tilde{\epsilon}^1}(D)$ *est une meilleure approximation de* $\Lambda_\epsilon(A)$ *que* $\Lambda_{\tilde{\epsilon}^2}(D)$». On matérialise donc la notion de «$\Lambda_{\tilde{\epsilon}}(D)$ *est une bonne approximation de* $\Lambda_\epsilon(A)$» par un nombre γ calculé comme suit. Supposons que $\tilde{\epsilon} = \alpha\epsilon$. Soit M_A et M_B les matrices de termes génériques respectifs $\log_{10}\left(\frac{\sigma_{min}(A - z_{kw}I)}{\|A\|}\right)$ et $\log_{10}\left(\frac{\sigma_{min}(D - z_{kw}I)}{\alpha\|D\|}\right)$. On définit γ par $\gamma = \|M_A - M_D\|$. Si $S = I$ alors $\kappa(S) = 1$ et l'approximation optimum est obtenue pour

$$\tilde{\epsilon} = \epsilon \quad \Rightarrow \quad \alpha = 1 \quad \Rightarrow \quad \gamma = 0 \ .$$

Plus γ sera proche de 0, meilleure sera l'approximation de $\Lambda_\epsilon(A)$. Dans un premier temps nous allons étudier pour une décomposition fixée de la matrice A de la forme (8), l'évolution de la valeur de γ en fonction du paramètre $\alpha = \frac{\tilde{\epsilon}}{\epsilon}$.

Cas où A est la matrice de GRCAR.

– Décomposition en $q = 2$ blocs

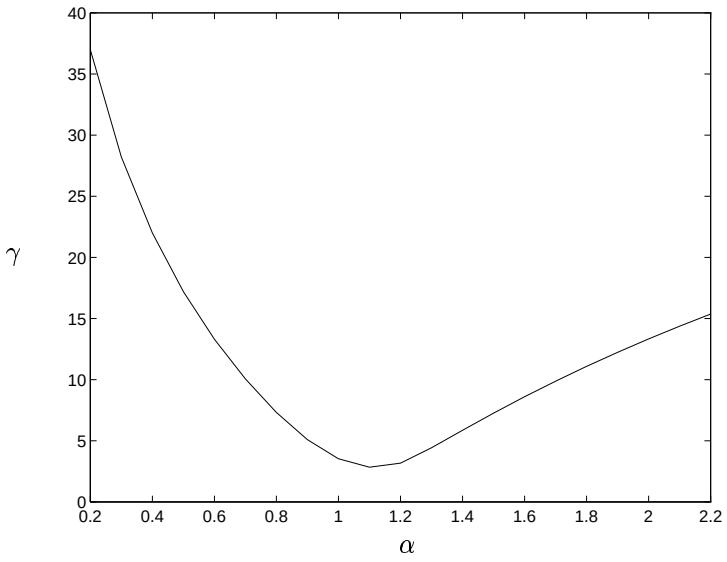

FIG. 36: *Évolution de γ en fonction de α pour une décomposition en 2 blocs de la matrice de GRCAR*

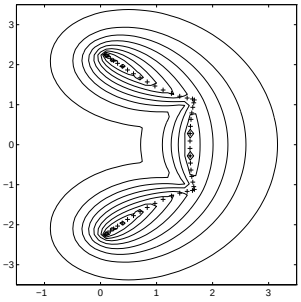

FIG. 37: *Portrait spectral de A*

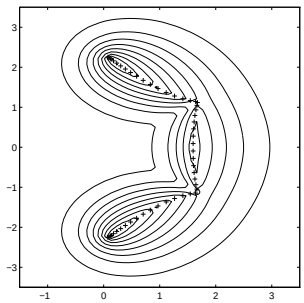

FIG. 38: $\alpha = 0.6$

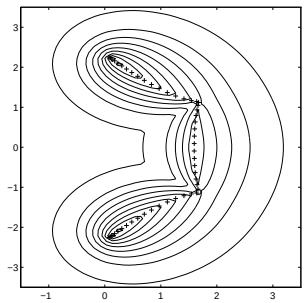

FIG. 39: $\alpha = 1.1$

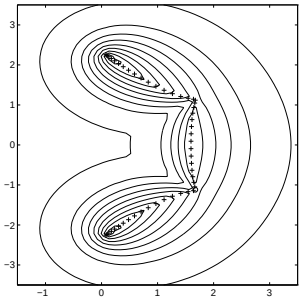

FIG. 40: $\alpha = 1.6$

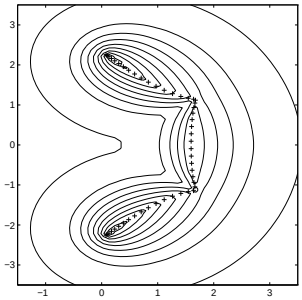

FIG. 41: $\alpha = 2.1$

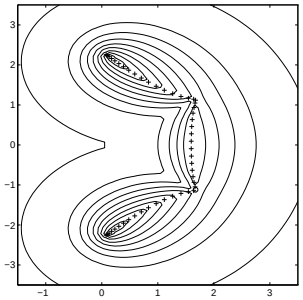

FIG. 42: $\alpha = 2.6$

– Décomposition en $q = 3$ blocs

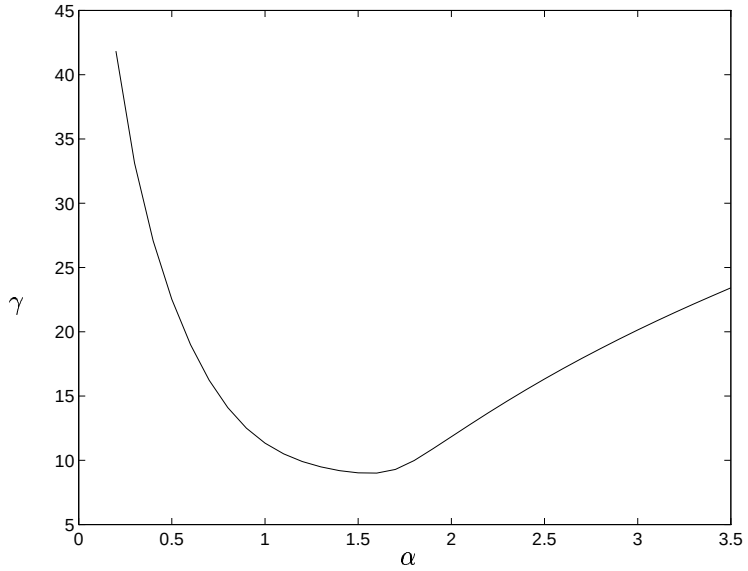

FIG. 43: *Évolution de γ en fonction de α pour une décomposition en 3 blocs de la matrice de GRCAR*

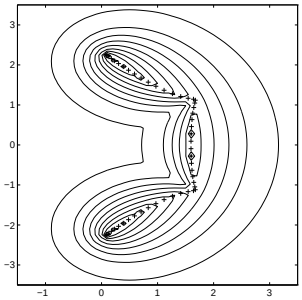

FIG. 44: *Portrait spectral de A*

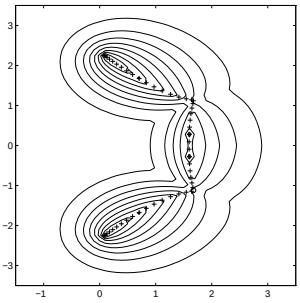

FIG. 45: $\alpha = 0.5$

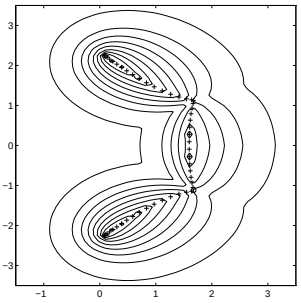

FIG. 46: $\alpha = 1$

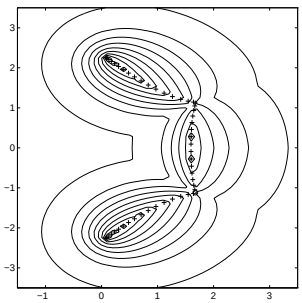

FIG. 47: $\alpha = 1.5$

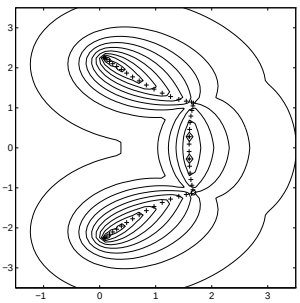

FIG. 48: $\alpha = 2$

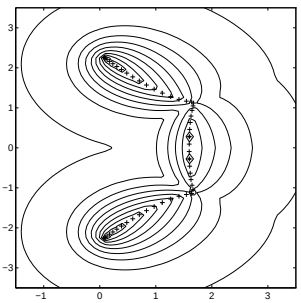

FIG. 49: $\alpha = 2.5$

Approximation de $\Lambda_\epsilon(A)$

Voyons maintenant quelques exemples d'approximation de $\Lambda_\epsilon(A)$ par $\Lambda_{\tilde{\epsilon}}(D)$. N'ayant pas trouvé de moyen de calculer la valeur $\tilde{\epsilon}_{Opt}$, on décide d'approcher $\Lambda_\epsilon(A)$ par $\Lambda_\epsilon(D)$ (i.e. de prendre $\tilde{\epsilon} = \epsilon$). On visualise l'évolution de cette approximation en fonction du nombre q de blocs de la matrice D résultant de la bloc-diagonalisation de A. Ainsi pour chaque décomposition, on visualise la structure creuse de la matrice D, le champ des valeurs de chaque matrice D_i et enfin $\Lambda_\epsilon(D)$. Les deux matrices A étudiées seront celle de GRCAR et celle de FRANK.

Exemple: matrice de GRCAR

Ordre de $A = 50$
V=matrice des vecteurs propres de A
$\kappa(V) = 2.0e+8$

Nombre de blocs	$\kappa(S)$
13	6188
12	5487
11	3903
10	3409
9	2167
8	1976
7	1684
6	1469
5	1372
4	1359
3	702.8
2	308.5
1	1.0

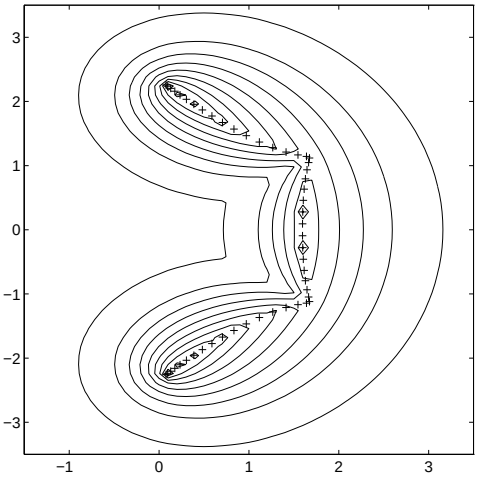

FIG. 50: *Portrait spectral de A*

- **Décomposition en $q = 3$ blocs :** $\kappa(S) = 702.8$

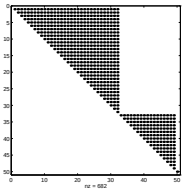

FIG. 51: *Stucture creuse de D*

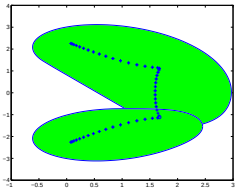

FIG. 52: *Champ des valeurs de D_i*

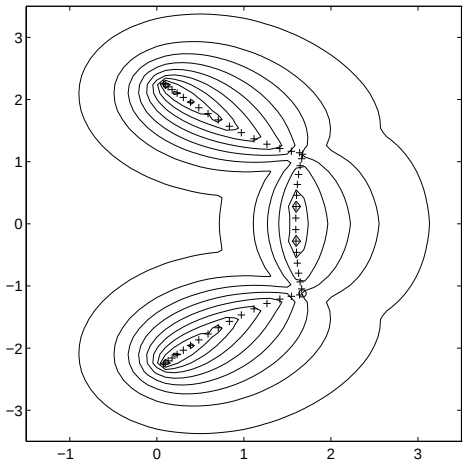

FIG. 53: *Portrait spectral de D*

- **Décomposition en $q = 6$ blocs :** $\kappa(S) = 1469$

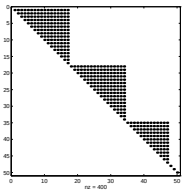

FIG. 54: *Structure creuse de D*

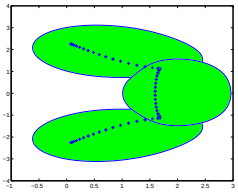

FIG. 55: *Champ des valeurs de D_i*

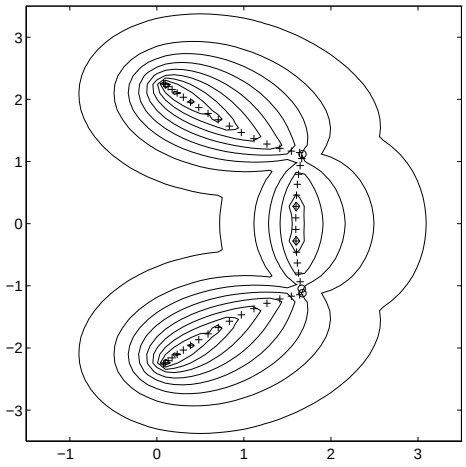

FIG. 56: *Portrait spectral de D*

- **Décomposition en $q = 13$ blocs :** $\kappa(S) = 6188$

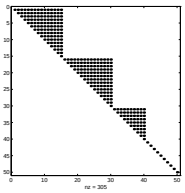

FIG. 57: *Structure creuse de D*

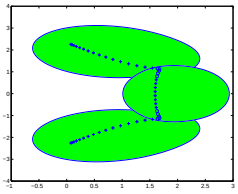

FIG. 58: *Champ des valeurs de D_i*

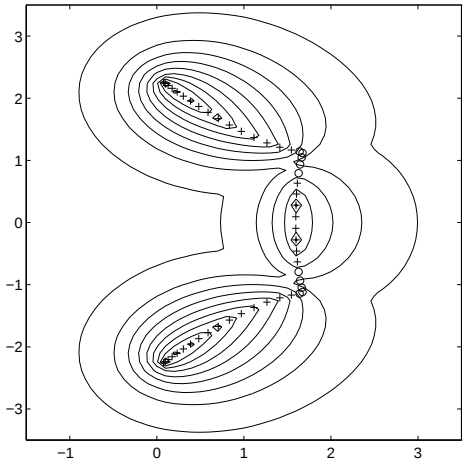

FIG. 59: *Portrait spectral de D*

Exemple: matrice de FRANK

Ordre de $A = 50$
V=matrice des vecteurs propres de A
$\kappa(V) = 5.55e + 10$

Nombre de blocs	$\kappa(S)$
10	9.784e+7
9	9.876e+5
8	2.096e+4
7	2615
6	2626
5	667.8
4	199.2
3	186.8
2	47.81
1	1.0

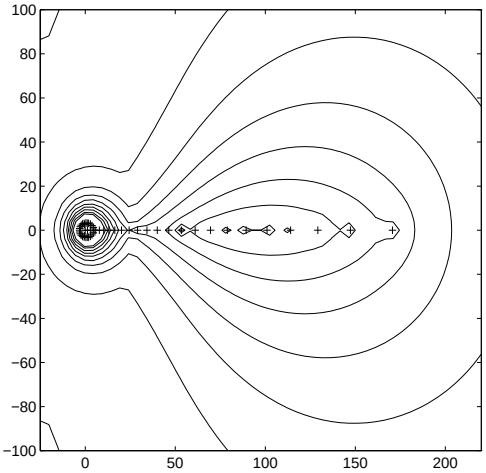

FIG. 60: *Portrait spectral de A*

- **Décomposition en $q = 2$ blocs :** $\kappa(S) = 47.81$

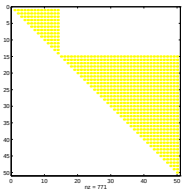

FIG. 61: *Stucture creuse de D*

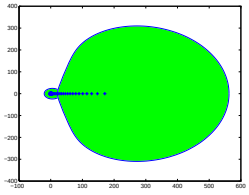

FIG. 62: *Champ des valeurs de D_i*

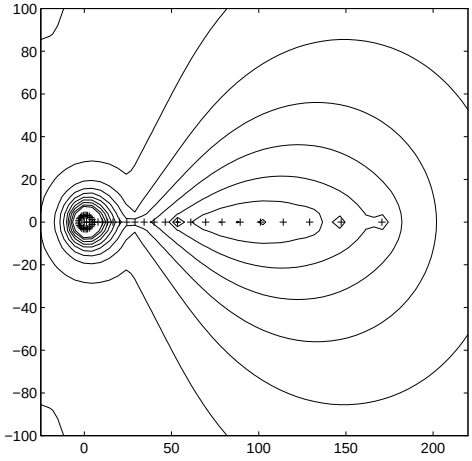

FIG. 63: *Portrait spectral de D*

- **Décomposition en $q = 6$ blocs :** $\kappa(S) = 2626$

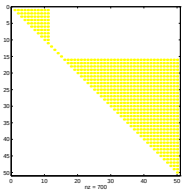

FIG. 64: *Structure creuse de D*

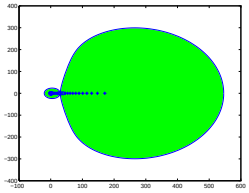

FIG. 65: *Champ des valeurs de D_i*

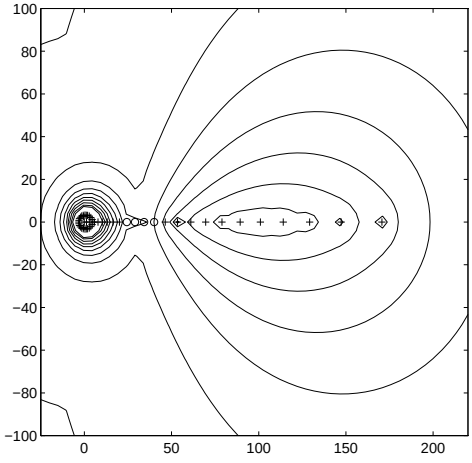

FIG. 66: *Portrait spectral de D*

- **Décomposition en $q = 10$ blocs :** $\kappa(S) = 9.784e + 7$

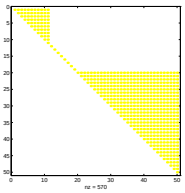

FIG. 67: *Structure creuse de D*

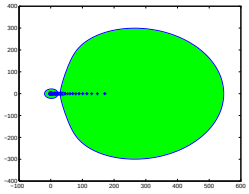

FIG. 68: *Champ des valeurs de D_i*

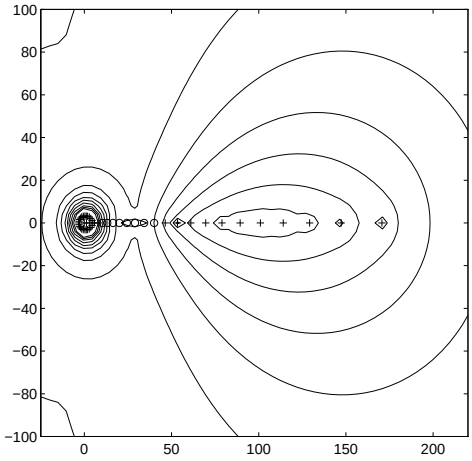

FIG. 69: *Portrait spectral de D*

- Dans ce deuxième exemple, on va s'intéresser plus spécifiquement au comportement de $\Lambda_\epsilon(D)$ lorsque le nombre q de blocs de D augmente. La matrice initiale A étudiée est celle de GRCAR. Dans cet exemple, on fait varier le nombre de blocs de 1 à 11, et pour chaque décomposition, on visualise le portrait spectral de D c'est à dire $\Lambda_\epsilon(D)$.

Matrice de GRCAR

$n = 50$

$\kappa(V) = 2.00e+08$

Nombre de blocs	$\kappa(X)$
11	3903
10	3409
9	2167
8	1976
7	1684
6	1469
5	1372
4	1359
3	702.8
2	308.5
1	1.0

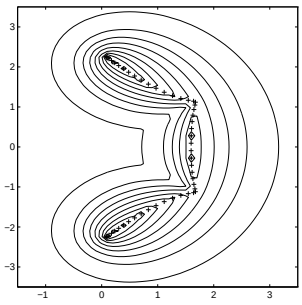

FIG. 70: *Portrait spectral de A*

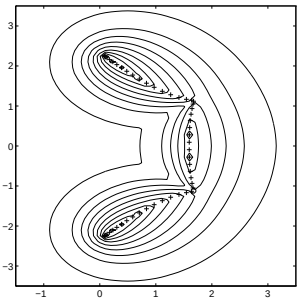

FIG. 71: *q=2* $\kappa(S) = 308.5$

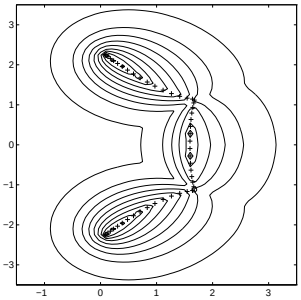

FIG. 72: $q = 3$ $\kappa(S) = 702.8$

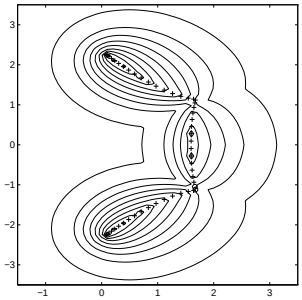

FIG. 73: *q=4* $\kappa(S) = 1359$

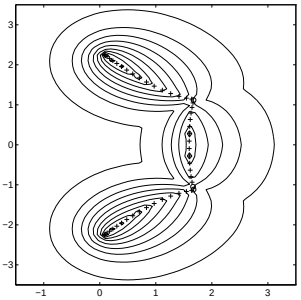

FIG. 74: *q=5* $\kappa(S) = 1372$

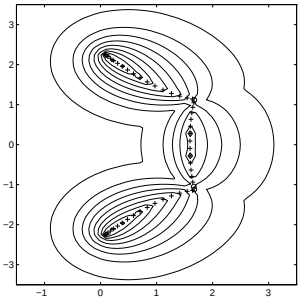

FIG. 75: *q=6* $\kappa(S) = 1469$

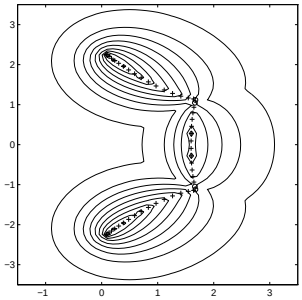

FIG. 76: *q=7* $\kappa(S) = 1684$

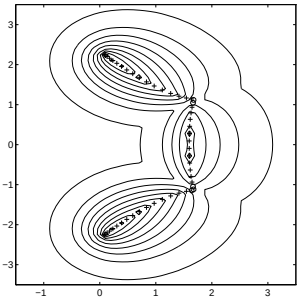

FIG. 77: *q=8* $\kappa(S) = 1976$

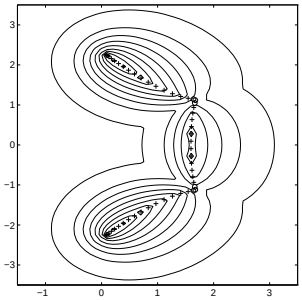

FIG. 78: *q=9* $\kappa(S) = 2167$

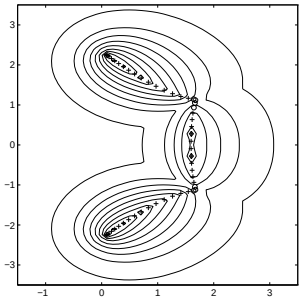

FIG. 79: *q=10* $\kappa(S) = 3409$

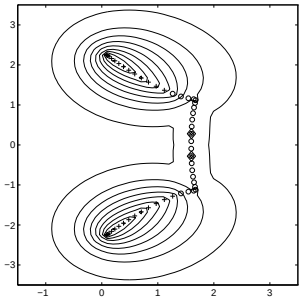

FIG. 80: *q=25* $\kappa(S) = 9.06e{+}4$

5 Optimisation

5.1 Remarque concernant le parallélisme

Comme on l'a remarqué au paragraphe 4.3, le calcul des différents $\Lambda_\epsilon(D_i),\ i = 1, \ldots, q$ étant totalement indépendant les uns des autres, il peut s'effectuer en parallèle. Cependant, dans l'optique d'obtenir une bonne répartition de charge de calcul, il est nécessaire que les blocs D_i soient de dimension comparable. En conséquence, si on désire utiliser efficacement ce parallélisme, il faut revoir la procédure de détermination des blocs de façon à trouver un compromis entre le fait d'avoir un conditionnement $\kappa(S)$ petit et des blocs de dimension comparables.

5.2 Réduction de la taille de la grille

Soit G la grille de $N \times N$ points, discrétisation d'une partie du plan complexe $\mathbb{C}$, qui sert pour le calcul du portrait spectral de la matrice A. Si on suit la démarche décrite au paragraphe 4.2, pour chaque point z_{kw} de cette grille, on calcule $\sigma_{min}(D_i - z_{kw} I_i),\ i = 1, \ldots, q$, ce qui nous permet de calculer $M_i(k,w)$ (i.e. élément en position (k,w) de la matrice M_i du paragraphe 4.2) et ainsi $M(k,w) = \min_{i=1,\ldots,q} (M_i(k,w)),\ 1 \leq k \leq N,\ 1 \leq w \leq N$. On peut se demander s'il est nécessaire de calculer tout les éléments de la matrice M_i? En d'autres termes, doit-on calculer $\sigma_{min}(D_i - z_{kw} I_i),\ i = 1, \ldots, q$ pour la totalité des $N \times N$ points de la grille G? Supposons par exemple que l'on sache déterminer une "sous-grille" G_{i_0} de G tel que

$$\forall z \in (G \backslash G_{i_0}) \quad \sigma_{min}(D_{i_0} - z I_{i_0}) > \min_{i=1,\ldots,q} \sigma_{min}(D_i - z I_i)$$

(i.e. en tout point extérieur à G_{i_0}, le minimum des σ_{min} est atteint pour un bloc $i \neq i_0$). Calculer $\sigma_{min}(D_{i_0} - z I_{i_0})$ pour z extérieur à G_{i_0} est alors inutile, puisque l'on sait par avance qu'il ne sera pas le minimum recherché. Ainsi on peut réduire le calcul de $\sigma_{min}(D_{i_0} - z I_{i_0})$ au seul $z \in G_{i_0}$.

Étant donnée une décomposition de A du type (8), toute la difficulté va consister à déterminer à priori les sous-grilles G_i sur lesquelles on va effectivement effectuer le calcul de $\sigma_{min}(D_i - zI_i)$ et tel que $G = \bigcup_{i=1}^{q} G_i$ (i.e. les sous-grilles G_i doivent être un recouvrement de G). Ce point n'a pour l'instant pas été approfondi. Le problème inverse est tout aussi intéréssant. Étant donné G_1 une sous-grille de G, cherchons une décomposition de A du type (8) tel que

$$\forall z \in G_1 \quad \min_{i=1,\cdots,q} (\sigma_{min}(D_i - zI_i) = \sigma_{min}(D_1 - zI_1))$$

(i.e. c'est le premier bloc qui réalise le minimum recherché). C'est l'objet du paragraphe suivant.

5.3 Cas où l'on ne s'intéresse qu'à une partie du portrait spectral

5.3.1 Méthodologie

Dans la plupart des cas, on ne s'intéresse pas à la totalité du portrait spectral, mais seulement à une partie de celui-ci. Par exemple, lorsque l'on travaille sur des problèmes de stabilité, seule la zone proche de l'axe des imaginaires est intéressante. Le reste du portrait spectral importe peu. Est-il possible alors d'optimiser notre procédure d'approximation, de façon à être le plus précis possible dans la zone intéressante, quitte à l'être beaucoup moins à l'extérieure de cette zone?
Bien sûr, et pour se faire, nous allons changer quelque peu notre procédure de bloc-diagonalisation de façon à regrouper dans un même bloc les valeurs propres qui se situent dans la zone $\mathcal{I}$ intéressante. Soit $(\lambda_i, u_i) \quad i = 1, \ldots, l$ les couples propres dont la valeur propre λ_i se situe à l'intérieur de $\mathcal{I}$. Ensuite, on recherche les valeurs propres $\lambda_i \quad i = l+1, \ldots, m$ susceptibles, de jouer un rôle de part leur proximité. Enfin, on recherche les valeurs propres susceptibles si elles n'étaient pas incorporées dans ce premier bloc, d'engendrer un «grand» conditionnement $\kappa(S)$ dans la décomposition (8). C'est à dire que l'on recherche les valeurs propres qui correspondent à des vecteurs propres dont le cosinus avec le sous-espace $\mathcal{E} = \mathrm{eng}(u_1, \ldots, u_m)$ est proche de 1. Pour plus de précision concernant cette détermination, se reporter au paragraphe 2.1.1, on procède de la même manière. Soit $(\lambda_i, u_i) \quad i = m+1, \ldots, p$ les couples propres ainsi déterminés. Le premier bloc est alors formé des valeurs propres $\lambda_i \quad i = 1, \ldots, p$. Un second bloc est formé des valeurs propres appartenant au complémentaire du spectre (i.e. $\lambda_i \quad i = p+1, \ldots, n$). Deux exemples de décomposition type sont donnés en figure (81) et figure (82). On visualise ainsi non seulement la répartition des valeurs propres entre les deux blocs, mais aussi la raison qui fait qu'une valeur propre appartient ou non au premier bloc. Les valeurs propres représentées par des "+" sont celles qui sont intérieures au domaine $\mathcal{I}$, celles représentées par des "$\times$" (respectivement des "$*$") sont des valeurs qui ont été incorporées au premier bloc de part leur proximité (respectivement de part leur cosinus proche de 1), enfin celles représentées par "." sont celles du second bloc.

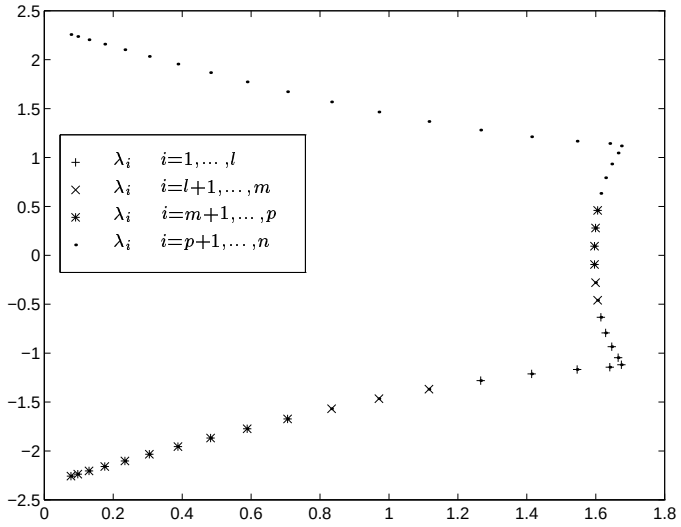

FIG. 81: *Répartition des éléments propres des deux blocs, matrice de GRCAR,* $\mathcal{I} = [1.2, 2] \times [-1.6, -0.6]$

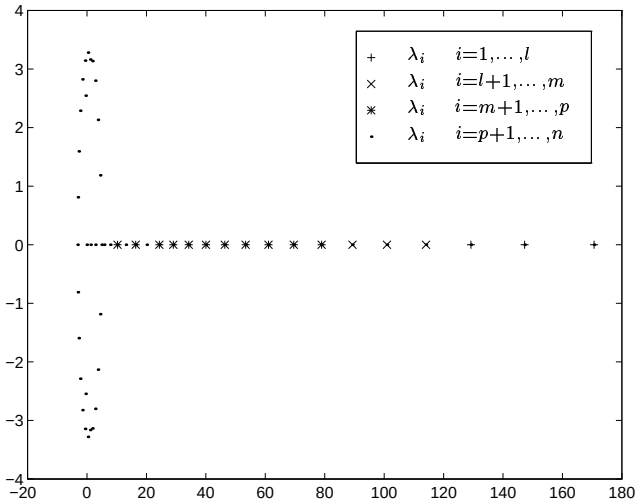

FIG. 82: *Répartition des éléments propres des deux blocs, matrice de Frank,* $\mathcal{I} = [120, 200] \times [-50, 50]$

On utilise notre procédure de réordonnacement décrite au paragraphe 2.1.2, pour réordonner les valeurs propres sur la diagonale de la factorisation de Schur. Ainsi on obtient la factorisation de Schur suivante:

$$Q^*AQ = T = \begin{pmatrix} T_{11} & T_{12} \\ & T_{22} \end{pmatrix} \tag{61}$$

avec $\Lambda_0(T_{11}) = \{\lambda_1, \ldots, \lambda_p\}$ et $\Lambda_0(T_{22}) = \{\lambda_{p+1}, \ldots, \lambda_n\}$.
On utililse alors les résultats du théorème (4) pour éliminer le terme extradiagonal T_{12} de la matrice T (voir paragraphe 2.1.3). Enfin on effectue l'orthonormalisation des blocs colonnes de S (voir paragraphe 2.1.4). On obtient une factorisation finale du type:

$$S^*AS = T = \begin{pmatrix} D_{11} & 0 \\ 0 & D_{22} \end{pmatrix} \tag{62}$$

Le portrait spectral de la matrice A dans la zone $\mathcal{I}$ est alors approché par celui de la matrice D_{11} dans cette même zone. Comme $\dim(D_{11}) \ll \dim(A)$, celui-ci se calcule sans difficulté et rapidement en utilisant une méthode basée sur la décomposition en valeurs singulières.
Les résultats numériques montrent qu'effectivement, l'approximation dans la zone $\mathcal{I}$ est relativement précise. Le coût d'une telle approximation, pour une grille discrétisée en $N \times N$ points est de l'ordre de $O(N^2p^3)$ plus le coût d'une bloc-factorisation de A. Ceci est à comparer au coût de calcul de $\Lambda_\epsilon(A)$ sur cette même grille qui de l'ordre de $O(N^2n^3)$, avec $p \ll n$.

5.3.2 Exemples numériques

Pour cet exemple (figure (83) et figure (84)), on prend pour A la matrice GRCAR d'ordre 50. On approche, pour différentes zones $\mathcal{I}$ le portrait spectral de A en trait plein, par celui de D_{11} en trait pointillé. On fait de même en figure (85) et (86) avec respectivement comme matrice A, la matrice de Pentoep et celle de Frank d'ordre 50.

– Matrice de GRCAR, $\mathcal{I} = [-0.5, 0.5] \times [-3, -1.5]$. On visualise en trait plein $\Lambda_\epsilon(A)$ et en pointillé $\Lambda_\epsilon(D_1)$. Dans cet exemple, on a $n = 50$ et $p = 17$.

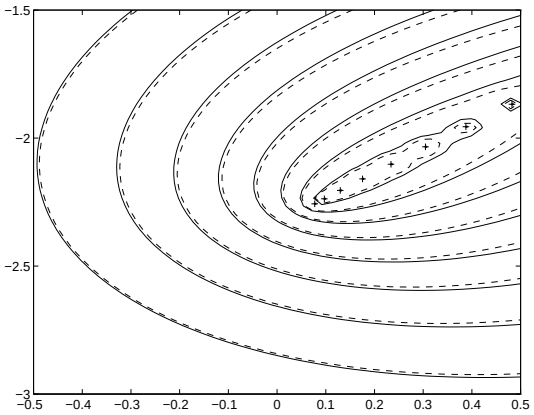

FIG. 83: *Comparaison de* $\Lambda_\epsilon(A)$ *et de* $\Lambda_\epsilon(D_1)$

– Matrice de GRCAR, $\mathcal{I} = [1.2, 2] \times [-1.6, -0.6]$. On visualise en trait plein $\Lambda_\epsilon(A)$ et en pointillé $\Lambda_{\epsilon Opt}(D_1)$.

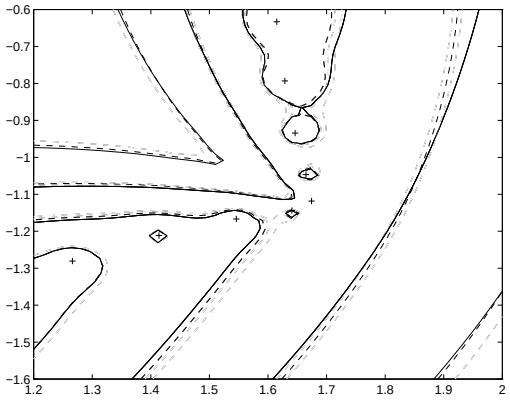

FIG. 84: *Comparaison de* $\Lambda_\epsilon(A)$ *et de* $\Lambda_{\epsilon Opt}(D_1)$

– Matrice de PENTOEP, $\mathcal{I} = [1, 1.6] \times [-0.4, 0.4]$.On visualise en trait plein $\Lambda_\epsilon(A)$ et en pointillé $\Lambda_{\epsilon Opt}(D_1)$.

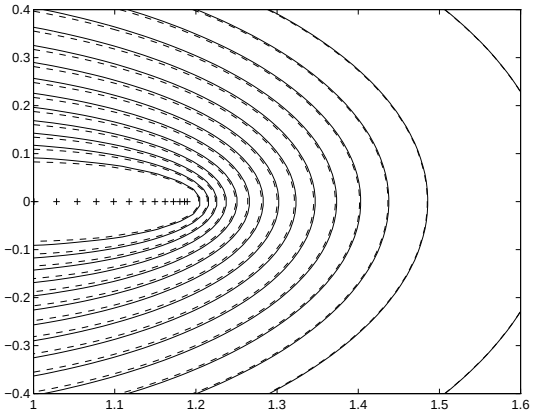

FIG. 85: *Comparaison de* $\Lambda_\epsilon(A)$ *et de* $\Lambda_\epsilon(D_1)$

– Matrice de Frank, $\mathcal{I} = [120, 200] \times [-50, 50]$. On visualise en trait plein $\Lambda_\epsilon(A)$ et en pointillé $\Lambda_\epsilon(D_1)$. Dans cet exemple, on a $n = 50$ et $p = 17$.

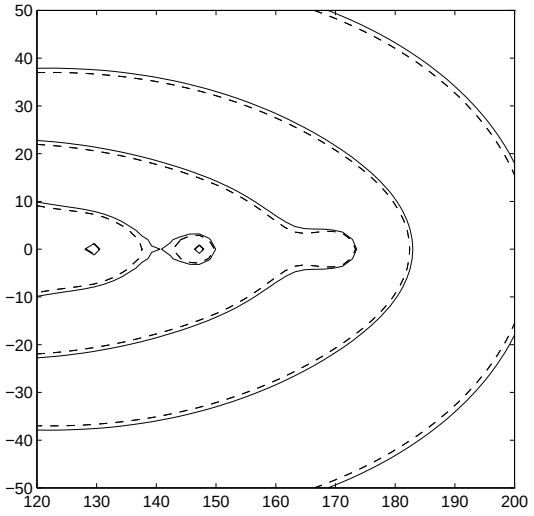

FIG. 86: *Comparaison de* $\Lambda_\epsilon(A)$ *et de* $\Lambda_\epsilon(D_1)$

Ces résultats sont probants. L'approximation du portrait spectral de A par celui de D_{11} dans la zone $\mathcal{I}$ semble relativement précis. Notons que dans chaque exemple, on a approché le portrait spectral d'une matrice d'ordre $n = 50$ par celui d'une matrice d'ordre $p = 17$.

Références

[1] C. A. Bavely and G. W. Stewart. An algorithm for computing reducing subspaces by block diagonalisation. *SIAM J. Numer. Anal.*, 16(2):359–376, 1979.

[2] J. Demmel. The condition number of equivalence transformations that block diagonalize matrix pencils. *SIAM J. Numer. Anal.*, 20(3):599–610, 1983.

[3] J. W. Demmel. Computing stable eigendecompositions of matrices. *Lin. Alg. Applic.*, pages 163–193, 1986.

[4] F. R. Gantmacher. *Théorie des matrices*, volume 1,2. Dunod, 1966.

[5] S. K. Godunov. Spectral portrait of matrices and criteria of spectrum dichotomy. *In Computer arithmetic and enclosure methods.*, 1991.

[6] G. H. Golub and Ch. Van Loan. *Matrix computation.* The Johns Hopkins University Press, 2 edition, 1989.

[7] G. H. Golub and J. H. Wilkinson. Ill-conditioned eigensystems and the computation of the jordan canonical form. *SIAM Review*, 18(4):578–619, 1976.

[8] N. J. Higham. The test matrix toolbox for matlab (version 3.0). Technical report, Manchester Center for Computational Mathematics, 1995.

[9] R. A. Horn and C. R. Johnson. *Topics in matrix analysis.* Cambridge University Press, 1991.

[10] B. Kågström and Axel. R. An algorithm for numerical computation of the jordan normal form of a. *ACM Trans. Math. Soft.*, 6(3):398–419, 1979.

[11] P. Lascaux and R. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur.* Masson, 1986.

[12] G. W. Stewart and J. Sun. *Matrix perturbation theory.* Academic Press Inc., 1990.

[13] L. N. Trefethen. Pseudospectra of matrices. *In Numerical analysis*, 1991.

[14] J. Varah. On the separation of two matrices. *SIAM J. Numer. Anal.*, 16(2):216–222, 1979.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399